DOCUMENT RESUME

ED 190 665 TM 800 464

AUTHOR Taylor, Hugh: And Others
TITLE Selection and Use of Standardized Tests: A Resource Book for School and District Personnel.
INSTITUTION Victoria Univ. (British Columbia).
SPONS AGENCY British Columbia Dept. of Education, Victoria.
PUB DATE Aug 79
NOTE 71p.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Diagnosis: Elementary Secondary Education: Program Development: Resource Materials: Scoring: *Standardized Tests: *Testing: Testing Programs: Test Interpretation: Test Results: *Test Selection
IDENTIFIERS Test Reporting

ABSTRACT

A statement of the major issues surrounding the use of standardized tests in schools is provided, covering the following areas of concern: (1) the major types of standardized tests and how they differ from teacher-made tests: (2) the role of standardized tests in education: (3) administering/scoring standardized tests: (4) interpretation of standardized test scores: (5) reporting results of standardized tests: and (6) establishing a standardized testing program. Some practical guides to aid in specific problems of selecting and using tests are listed in the appendices: a checklist is cross-referenced to the rest of the book so items may be explored in detail: a glossary of measurement terms: a list of test review periodicals: a list of test publishers: descriptions of selected tests: test administration guide: and, resources essential to a district testing program. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190665

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Selection and Use of Standardized Tests

A Resource Book for School and District Personnel

TM 800464

Province of British Columbia
Ministry of Education, Science and Technology
LEARNING ASSESSMENT BRANCH

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
P. Northover
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SELECTION AND USE OF STANDARDIZED TESTS:

A RESOURCE BOOK FOR SCHOOL AND DISTRICT PERSONNEL

Prepared by

Hugh Taylor, University of Victoria
Robert Wilson, Learning Assessment Branch
R. Nancy Greer, Learning Assessment Branch

Learning Assessment Branch
Ministry of Education, Science and Technology
Province of British Columbia

August, 1979

JUN 2 0 1980

PREFACE

It is a fairly recent phenomenon in education to expect that decisions affecting children and the programs designed for them and society should be buttressed by solid, trustworthy information. The area of standardized testing has long stood for the provision of reliable information to aid in decision-making. Often, the data accumulated by formal testing has been misused or ignored in planning for students' futures and program development, but the overall record has been promising. To further this promise, and to guard against misapplication and nonuse, this resource book has been developed.

The book has two major purposes. First, it provides a relatively compact statement of the major issues surrounding the use of standardized tests in the schools. Secondly, it lists in its appendices some practical guides to aid in specific problems of selecting and using tests. This handbook, then, can be read through as background material for those interested in the basic issues or it can be used as a guidebook for specific problems in test selection, purchase, and review. For example, readers involved in a specific selection decision should refer to the procedures checklist included as Appendix F. It is cross-referenced to the rest of the handbook so that each item can be explored in more detail. Other appendices provide a glossary of terms, a list of periodicals that review tests, a list of publishers, a description of some commonly administered tests, a guide for administration, and a short list of resources essential to a district testing program.

Appreciation is expressed here to Dr. Hugh Taylor of the University of Victoria for his contributions to this manual during its preparation stages. The coverage of the important issues reflected in his early drafts, and his advice throughout the project, are gratefully acknowledged.

R. Nancy Greer, Director
Learning Assessment Branch
Ministry of Education,
Science and Technology

August, 1979

TABLE OF CONTENTS

Chapter		Page
1.	INTRODUCTION	
	1.1 What are Standardized Tests?	8
	1.2 How do Standardized Tests Differ from Teacher-Made Tests?	8
	1.3 What are the Major Sources of Information About Standardized Tests?	10
	1.4 What are the Major Types of Standardized Tests Used in Schools?	10
2.	THE ROLE OF STANDARDIZED TESTS IN EDUCATION	
	2.1 What Purposes can be Served by Information from Standardized Tests?	13
	(a) Selection	
	(b) Placement	
	(c) Diagnosis and Remediation	
	(d) Program Planning and Evaluation	
	2.2 What Guidelines Should be Applied when Selecting a Standardized Test?	20
	(a) Validity	
	(b) Reliability	
	(c) Adequacy of Norm-Referencing	
3.	ADMINISTERING AND SCORING STANDARDIZED TESTS	
	3.1 When Should Standardized Tests be Administered During the School Year?	23
	3.2 How Should Standardized Tests be Administered and Who Should do it?	24
	3.3 How Should Standardized Tests be Scored and Who Should do it?	24
4.	INTERPRETATION OF STANDARDIZED TEST SCORES	
	4.1 What are the Differences Between "Norms" and "Standards"?	27
	4.2 How can Local Norms be Established and Used?	28
	4.3 What is Measurement Error?	29
	4.4 Why is the Normal Curve Used as the Basis for Most Standardized Test Scores?	29
	4.5 How Should Scores be Interpreted?	32
	(a) Percentile Ranks	
	(b) Stanines	
	(c) Standard Scores - T-Scores, Intelligence Quotients	
	(d) Grade Equivalent Scores	

TABLE OF CONTENTS CONTINUED

Chapter		Page
5.	REPORTING RESULTS OF STANDARDIZED TESTS	
5.1	How can Misinterpretation be Avoided?	39
5.2	What are Some Alternative Reporting Formats?	39
5.3	How Should Information Best be Passed on to People?	40
6.	ESTABLISHING A STANDARDIZED TESTING PROGRAM	
6.1	What are the Essential and Optional Components of a Testing Program?	43
6.2	What Steps are Involved in Planning a School Testing Program?	44
6.3	How Should Tests be Selected?	45
6.4	What Procedures Should be Followed to Evaluate a Testing Program?	46
REFERENCES		51
APPENDICES		
A.	A Glossary of Measurement Terms	53
B.	Periodicals That Review Standardized Tests	63
C.	A Selective List of Test Publishers	65
D.	A Selected List of Standardized Scholastic Aptitude, Achievement, Interest Tests and Instruments	67
E.	Test Administration Guide	75
F.	Selecting and Using a Standardized Test: Procedures Checklist	77
G.	District Level Resources Essential to a Viable Testing Program	81

CHAPTER 1

INTRODUCTION

Educators are constantly faced with the responsibility of making decisions which directly or indirectly affect the lives of children. Often, the information available to aid in the decision-making is incomplete, subjective, and fraught with the possibility of error. Should Mary be taken from the regular classroom and given specialized help? Should a more phonics-based reading program be purchased or developed to help the low achievers in our district learn to read more effectively? Should Stephen be encouraged to pursue university level training in his field of interest, engineering? The human and social costs of leaving Mary in the regular stream when she needs special help (or, indeed, providing her with special help when she doesn't need it); of maintaining an inappropriate reading program when a change would be beneficial; of encouraging Stephen to attend university when he will drop out after his first year--all of these are dramatic choices yet represent the kind of decisions we must make continuously at all levels in the educational enterprise.

One thing, however, is clear. The more reliable and valid the information we have, the more likely we are to make correct choices for individuals and programs. No single source of information will provide all the information we need about a decision. A decision affecting Mary could not be made without eliciting information from her teacher, her parents, and the specialists to whom she might be entrusted. Their viewpoints are crucial, but not necessarily conclusive. None of these would be able to provide information about questions like: Given that Mary is having difficulty in her present class, how does she rank in her learning compared to other children of her age beyond her own class and school boundaries? Or, is Mary's achievement of such inadequacy that she fits a group of children found in other places to benefit from specialized training? The appropriate source to turn to for the answers to these questions would be one or more standardized tests selected and used with Mary so that the people charged with her care can have access to information.

Standardized instruments have some highly visible advantages over more informal educational measures. They are objective both in source and in design. They "look" impressive, not only in format, but because a large technology is responsible for their construction, adminstration, and scoring. Their long history has invested them with authority that experience over time has generally solidified. Their use in identifying the Marys who need special help, in aiding individuals in career and educational decisions, in providing information about the effectiveness of programs, and in diagnosing weak spots in curriculum and instruction has been well documented.

Some of these very strengths, however, can be counted as weaknesses when these instruments are asked to perform tasks for which they are not suited. Few group standardized instruments, for example, provide precise data for teachers

to use to diagnose specific deficiencies in individual pupil achievement. Also, the very reputation of the tests may force a too heavy reliance on this single source of data for such decisions as assigning students to "tracks" or altering program content. Finally, the tests and their results have too often been used for formal or informal assessment of teaching competence, a goal which may be worthy but which certainly can not be attained solely by the survey testing of students.

This resource book is largely concerned with providing you with information about the legitimate uses of standardized tests. In knowledgeable hands, they have important roles to play in educational decision-making. The book will also provide you with the background to select and use these tests wisely. If it achieves these twin purposes, it will fulfill its function well.

1.1 What are Standardized Tests?

Broadly defined, tests are "standardized" to the extent that they have been administered and scored under standard or uniform conditions. This standardization of procedures for administration and scoring then allows for comparisons to be made among students who have written the test at different times and in different locations. The basis for these comparisons are "norms", the distributions of scores attained by some specified group under uniform conditions of administration and scoring. More narrowly defined, the term "standardized" has become synonomous with "published". For the purposes of this booklet and to avoid any confusion, the term "standardized test" will be used here to refer to published tests that are accompanied with tables of norms. Further, the emphasis of this discussion will be on paper and pencil instruments which can be group-administered and which do not require the expertise of highly trained psychologists for their administration and interpretation.

1.2 How do Standardized Tests Differ from Teacher-Made Tests?

Standardized tests differ from teacher-made classroom tests in several ways.

(1) Because standardized tests are designed to be used in typical schools nationally as well as internationally, they are generally based on educational objectives that are widely shared by many school districts. As such, the content of standardized tests are usually much broader and related to more general academic skills than are teacher-made tests. In some situations the actual test booklet may contain a battery of four or more different subject-matter tests such as Word Knowledge, Reading Comprehension, Spelling, Mathematical Concepts, and Mathematics Problems. It thus becomes an important responsibility of the test user to determine the degree to which the each test's content reflects locally held curriculum emphases.

(2) Unlike the typical procedures used when teachers construct tests for their own use, the construction of standardized tests is far more complex and costly. Usually the test publishing company employs measurement specialists, curriculum designers, and subject

matter experts who work as a team. Current curricula are reviewed, a table of specifications is planned, test items are written and revised following extensive field testing and data analysis, items for different age groups are assembled into forms of the test, and norms are established through administration of the test to large samples of representative individuals for whom the test is being designed. Ideally, this developmental process will include thorough examination of the reliability and validity of the test for its intended uses, and the documentation of this information in a manual accompanying the test. Once again, it becomes the responsibility of the test users to critically examine this information in judging the appropriateness of the test for their purposes.

(3) Standardized tests, by definition, require absolute invariance in the procedures for administration and scoring. While it is unlikely that a teacher administering a self-developed classroom test would permit widely differing rules (i.e. allowing some students to use dictionairies while other may not), in standardized testing strict adherence to the procedures described in the test manual are critical to the valid interpretation of the results. Standardized tests are usually strictly timed, with directions for administration read by the teacher to younger children or written in detail for older students. Scoring is likewise an objective process in that no judgements are normally required of the scorer. Many standardized tests are answered on specially designed answer sheets that allow for automated machine scoring.

(4) Many of the more widely used standardized tests are produced in two or more different forms, unlike the usual "one-shot" nature of classroom tests. Forms contain different items but are based on the same specifications of content. In this way scores on the different forms can be compared during interpretation. Some standardized tests in content areas such as reading and mathematics are also produced in different grade level forms, and through statistical equating of their scores, can be used for measuring student progress over a period of a few years.

(5) Perhaps the most important difference between standardized tests and teacher made tests is in the interpretation of the test scores. Publishers of standardized tests provide manuals containing tables of norms based on the actual performance of representative groups of students. The norms equate raw scores into various types of derived scores. As such, norms provide a frame of reference for estimating an individual's or group's achievement level or for comparing relative levels of achievement on several different tests that were standardized on the same sample of students. Test developers know in advance that the measurement of differences is an intended use of test scores, and therefore they select items during the test development stage for inclusion on the final form of the test that give maximum statistical precision to these types of numerical comparisons. A wide range of student achievement, therefore, is a stated objective of these tests. Users must expect to find such a range, then, when standardized tests are administered to their students.

1.3 What are the Major Sources of Information About Standardized Tests?

The most important sources of information about standardized tests are the Mental Measurements Yearbook series (1936-1978), edited by Oscar K. Buros. The latest edition, the 8th published in1978, consists of two volumes listing over 12,000 references on specific tests. These yearbooks are invaluable references for schools or districts planning to start a standardized testing program or wishing to conduct an evaluation of their current program. Along with providing a comprehensive listing of tests available, they provide valuable information on such questions as: How, when and on whom was the test normed? What evidence exists as to the reliability of the test? What do the test developers say concerning the valid use of the test? What evidence is there of its validity for other uses? Does the latest review concur with the claims of the test developers? What cautions and recommendations are given to the user of the test by the developer and the reviewer? The MMY is generally considered the best and most authoritative resource for objective information about any specific test and tests in general.

Buros also publishes other measurement references including Tests in Print II (1974) which contains a listing of all current tests, their publishers, costs, etc. and, if applicable, the actual Mental Measurements Yearbook reference that contains a review of each test. A catalogue of publications edited by O.K. Buros may be requested from:

Gryphon Press
220 Montgomery Street
Highland Park, N.J.
80904

For information specific to the standardized testing of reading, the International Reading Association has produced several publications which provide critical reviews of available tests. Appendix B contains a listing of these publications along with other periodicals that frequently review standardized tests. Appendix C lists the major test publishing companies. Publishers issue yearly catalogues from which specimen sets of tests may be ordered at a minimum cost. Such sets usually contain one form of a test, a scoring key, answer sheets, an administrator's manual, and occasionally a technical manual. Before a given standardized test is purchased for use, an examination of a specimen set must be made so that the appropriateness of the test's content, norms, testing time etc. can be assessed.

1.4 What are the Major Types of Standardized Tests Used in Schools?

For anyone who has examined the publications listed above, it becomes immediately obvious that a myriad of standardized tests numbering in the thousands are available for use. On closer examination, it will be seen that these fall into a number of categories, including tests of:

(1) aptitude or ability (both general and specific areas such as mechanical, clerical, etc.);

(2) achievement (both general and specific curriculum

areas);

(3) attitudes (towards school, home, work, etc.);

(4) personality, and

(5) interest (academic, vocational, extracurricular, etc.).

The two principal types of tests used rather extensively in schools are achievement tests and scholastic aptitude tests. A superficial study of an achievement and a scholastic aptitude test would reveal many similarities and only a few obvious differences. However, the tests do differ along a number of important dimensions:

Differences between achievement and scholastic aptitude tests.

(1) One important difference is the breadth of the content covered by the test. Achievement tests, particularly those in a specific subject matter area such as mathematics or reading are rather limited in content but do cover those subjects taught formally in schools. Scholastic aptitude tests, such as the Canadian Lorge-Thorndike, on the other hand, often contain items measuring verbal facility and quantitative reasoning. These abilities represent general but more informal aspects of development not specifically addressed by a school's curriculum.

(2) A second dimension of difference is the extent to which items on a test are defined by specific curricular programs. An achievement test such as the Canadian Test of Basic Skills is designed so that the sub-tests measure a limited academic area. A scholastic aptitude test, by comparison, samples learning in school as well as in the broader social culture. This difference requires that achievement tests be revised whenever the curriculum changes whereas scholastic aptitude tests are more independent of the content changes in schools over short periods of time.

(3) A third dimension of difference is the recency of the learning sampled by the test. An achievement test usually measures learning that has taken place over the previous year or so while an aptitude test may measure learning extending over a number of years.

(4) The fourth and major distinction between an achievement test and a scholastic aptitude test concerns the use of the test results. An achievement test and its sub-tests attempt to measure a student's current level of learning in particular academic areas. As such, the user should be concerned mainly with the content validity of the test for a certain classroom or grade level. Scholastic aptitude tests are used primarily to predict an individual's future school performance and thus the concept of criterion-related validity (or predictive validity) takes precedence when interpreting the test results. (A practical procedure for judging the criterion related validity of a test is discussed in Section 2).

A selected list of commonly used aptitude and achievement tests is provided in Appendix D.

CHAPTER 2

THE ROLE OF STANDARDIZED TESTS IN EDUCATION*

2.1. What Purposes can be Served by Information From Standardized Tests?

Standardized tests are administered with the expectation that measurements derived from them will be helpful in making better decisions than could be made without them. Although every decision involves the element of risk, not all risks lead to serious consequences. Because standardized tests are relatively expensive to purchase, administer, score and interpret, they derive their greatest benefit when the decision to be made is important and when information is not readily available from other less expensive sources.

The extent to which tests minimize risk and increase the probability of effective decisions is related to the amount of dependable knowledge the test provides beyond that which is already known. Using a test implies that there is a relationship between the measurements or scores derived from tests and the ability to make effective decisions. Having students take a spelling test presumes a relationship between the students' scores and the decision, for example, to recommend remedial work, additional exercises, or a revision of the curriculum. The following sections discuss four of the most frequent uses of standardized test results:

2.1.(a) Selection

The use of standardized test results to aid in deciding who will be accepted or rejected from an institution or program is far more common at the post-secondary level than during the elementary or secondary school years. Some of the principles involved are of sufficient importance, however, that they should be mentioned in this manual.

A test used for selection decisions must be proven to have high predictive or criterion-related validity. That is, the test must be able to identify those who will be successful from those who will not with the minimum of risk to the individual or the institution.

Unfortunately, tests are not infallible. All selection decisions, whether they be admission of an adult into medical school or the early entrance of a 4-year old into kindergarten, involve some risk of rejecting a potentially successful candidate or admitting one who subsequently fails. In order to minimize these risks, most selection procedures bring into play several sources of information rather than relying on one single test score. For example, universities typically consider senior high school grades, letters

**Substantial portions of this section of the manual have been taken from Sax, G., The Use of Standardized Tests in Evaluation., in Popham, W.J. (Ed.) Evaluation in Education: Current Applications. See References.*

of recommendation, and (in some cases) college entrance exam scores when making selection decisions. These multiple measures are then combined in a manner that has been determined empirically to minimize the risk of a wrong decision about a student. Statistically, what one strives for when designing a selection procedure is to maximize the correlation (degree of relationship) between scores on the selection battery and scores received later on measures of performance in the actual program for which the students are being selected.

Perhaps one of the simplest ways to evaluate the effectiveness of a selection procedure is to construct an expectancy table such as the one shown in Figure 2.1. In this hypothetical example, 100 children who were recommended by their teachers were given a battery of tests prior to being admitted to a special program for gifted and talented children. A selection committee made the tentative judgement that a combined score of 65 was indicative of minimally acceptable performance on the selection tests. However, in order to examine the validity of their judgements, they admitted all of the 100 recommended children during one academic year. At the end of the year, all children were rated for success in the program using a variety of measures that included teachers' evaluations and objective tests of achievement.

	Scores less than 65 Decision - Deny Admission	Scores 65 or better Decision - Admit	TOTAL
Were successful in the program	"False Negatives" 20 children	"True Positives" 52 children	72 children
Were unsuccessful in the program	"True Negatives" 16 children	"False Positives" 12 children	28 children
TOTALS	36 children	64 children	100 children

Figure 2.1

Expectancy Table: Selection of Children for a Special Program

From Figure 2.1, you will note that a total of 36 (20+16) children scored less than 65 on the selection battery and would normally have been denied admission. Sixteen of these 36 children did indeed have difficulties in the program (True Negatives), but note that 20 of those 36 children who would have been denied admission were in fact successful (False Negatives). On the right side of the table, it can be seen that 52 of the children who were admitted based on their score on the selection test were indeed successful (True Positives). Twelve of the children who did exceed the cutoff, however, were unsuccessful (False Positives). Thus, of the 64 children (52+12) who would have been admitted to the program based on the selection test, 52 would have been successful. This ration of 52/64 or 81% is referred to as the success ratio.

Of more importance, however, is the fact that a total of 32 children (20 False Negatives + 12 False Positives) of the original 100 were misclassified by

the selection test. The availability of this information would now allow the selection committee to examine the costs, human and otherwise, of such errors as they evaluate the procedures used for selection of children to the program. More importantly, it provides a valuable source of information for planning selection procedures for future years. With the addition of other information, the success ratio might be substantially improved.

2.1(b) Placement

Placement decisions in education address questions such as "What program of studies should a student be advised to pursue; what reading group should he be assigned to, and which special classes, (remedial, advanced, etc.) will he take?" A test that is useful for selection may not be effective for the more specific kinds of decisions that are involved in placement. Ideally, a placement test will reliably predict success in one program and at the same time discourage or reject placement in all other alternative programs.

Placement decisions fall into two general categories: vertical and horizontal placement. Vertical placement is required when it is necessary to estimate the grade or level within a grade at which a child is best able to function. The risk to the child in human terms of errors in placement decisions (for example, erroneously placing a young child with an undetected hearing problem in a class for slow learners) serves to emphasize the need for taking into account numerous other subjective and objective factors in addition to results of standardized tests. It goes without saying that when more radical placement decisions must be made that have long term consequences to the child's future as a student, they should not be undertaken without extensive individual testing of the child by personnel trained specifically for this task.

Day to day vertical placement decisions that a teacher may make, such as assigning children to one of several reading groups, can be aided by an expectancy table of the sort shown in Table 2.1. This table shows the relationship between a scholastic aptitude test administered in the Fall and a reading test administered in the Spring. It provides a means of estimating the criterion related or predictive validity of this aptitude test.

Local stanine* norms were based upon the test results of 437 students at the grade 5 level. The main body of the table shows the percentage of students at each stanine level on the predictor (scholastic aptitude) that received a particular stanine score on the criterion (Reading Test). In general, the table shows that students who score high on the ability test tend to score high on the reading test and vice versa, a fairly common finding.

See glossary for a definition of this derived score.

TABLE 2.1

Expectancy Table Showing Students' Stanines
on a Reading Test and a Test of School Ability

School District XYZ	DISTRICT NORMS									Grade V
Scholastic Aptitude Stanine	Percent Receiving Each Reading Stanine									Number of Students Within Each Scholastic Aptitude Stanine
	1	2	3	4	5	6	7	8	9	
9					5	21	21	16	37	19
8				10	14	28	24	10	14	29
7			4	12	22	26	26	8		49
6		4	11	15	27	20	17	5		74
5		4	13	25	30	23	5			89
4	5	11	30	21	21	11				76
3	6	29	27	20	10	8				51
2	18	18	41	12	9	3				34
1	31	12	25	12	19					16
										437

The Scholastic Aptitude Test was administered in October and the Reading Test was administered the following May.

An expectancy table such as this allows a teacher or counsellor to estimate the probability of success of a student at the beginning of the school year, provided the student's score on the aptitude test is available and provided the student can be considered typical of the students who took part in the local norming process. For example, a child entering grade 5 the next year could be assigned to a particular reading group with greater precision because of the known capability of the aptitude test's power of prediction.

A number of cautions are necessary in using expectancy table data. A small number of cases (such as one class) severely reduces the amount of confidence that can be placed in the predictions. Accumulating data over several years will help alleviate this problem, provided the characteristics of the student population and curriculum programs do not undergo exceptional changes. With a small number of students one can use fewer categories, such as using a 4-by-4 cell table instead of one which is 9-by-9. This will place a larger number of students in each cell and permit greater confidence in judgements based on the table. Expectancy tables are usually valid only for the school or the district in which they were developed. The data from one school may or may not be representative for another school.

Horizontal placement decisions involve determining which of several programs a student should pursue, as might be the case when a high school student selects a program of study from among a number of curricular options. Effective placement decisions of this sort are usually made on the basis of scores on several standardized tests. These would include an aptitude battery such as the Differential Aptitude Tests, as well as some assessment of the student's interests (eg. the Kuder General Interest Survey or the Strong Vocational Interest Blank). The distinction between aptitude and interest is important in assisting students to make vocational decisions. For example, an expressed interest in medicine does not specify the level at which the student would be most proficient - doctor, nurse, technician, orderly, practical nurse, etc. Aptitude scores when combined with results from other measures would assist in predicting the level at which the student is likely to be successful.

2.1(c) Diagnosis and Remediation

A test serves a diagnostic function when it is used to identify the strengths and weaknesses in the performance of individual students so that follow-up designed to improve performance can be undertaken. Diagnosis serves no end in itself but is a prerequisite step if reasonable treatment decisions are to be made. The teacher or counsellor is better able to recommend the specific skills in which the child is deficient.

As an initial screening for the identification of general areas of weakness, subtest scores on achievement batteries such as the Canadian Tests of Basic Skills or the Metropolitan Achievement Tests can be used. This initial screening can then be followed by more detailed diagnostic testing with weak students in the specific content areas identified as a problem. Tests designed for this type of follow-up, such as the Stanford Diagnostic Arithmetic Test, differ from the more general achievement batteries in that they have far more specific objectives and a larger number of items measuring each objective. These two differences allow for the reliable documentation of the specific nature of the child's problem that is not possible with general achievement batteries.

2.1(d) Program Planning and Evaluation

An essential feature of responsible educational management at all administrative levels (classroom, school, district and province) is the use of information about student performance to effect appropriate changes which will ensure that a greater percentage of students are successful in their schooling. Standardized test results can serve as valuable information in aiding decisions regarding programs. For example, group profiles of results from annually administered achievement batteries serve to document general patterns of strength and weakness. When coupled with data from classroom tests and provincial learning assessments, these may be instrumental in effecting needed program planning and revision.

When information is systematically and regularly collected, radical changes in performance across the students as a group can be documented, which can in turn lead to the identification of the conditions (positive and negative) responsible for this change.

The annual documentation of standardized test performance of the students can be of particular value when the effectiveness of a new program is being evaluated. For example, if district reading norms accumulated over 4 or 5 years are available, then the impact of introducing a new reading program can be assessed through comparisons of current group performance with that of previous years.

It should be emphasized here that there are potential problems associated with the use of standardized tests for program evaluation purposes. One such problem is a lack of congruence between what the standardized test measures and what the program is designed to achieve. As stated earlier in this booklet, standardized tests tend to reflect the most commonly held curricular goals and objectives as identified at the time the tests were developed. These objectives may not coincide with the objectives of an innovative program or a recent development in curriculum. If a program is to be given a "fair" evaluation, it is important that the test be congruent with the specific objectives of that program. In most cases standardized tests will need to be supplemented with other, perhaps locally developed, evaluations.

An additional caution is needed, particularly with regard to the use of standardized tests to evaluate programs intended for low achieving students. Because most standardized tests are multiple choice in format, there is a given probability that students will answer some questions correctly by guessing alone. This is usually referred to as the "chance score." On a 100-item test, with each question having four response options, the chance score = 100/4 =25%. Many low achieving students may be scoring well within the range of this "chance score", and it cannot therefore be assumed that any meaningful measurement has been made. In any situation where the students being assessed differ significantly from the students used to standardize the test, the information on reliability and measurement error presented by the test publisher should not be relied upon too heavily.

Before leaving the topic of program evaluation, two further points should be mentioned.. First, when the effects of a program on achievement are being evaluated, it is advisable to be on the look-out for "trait-treatment interactions". These are in evidence when the program is

more effective with some sub-groups of students than with others.

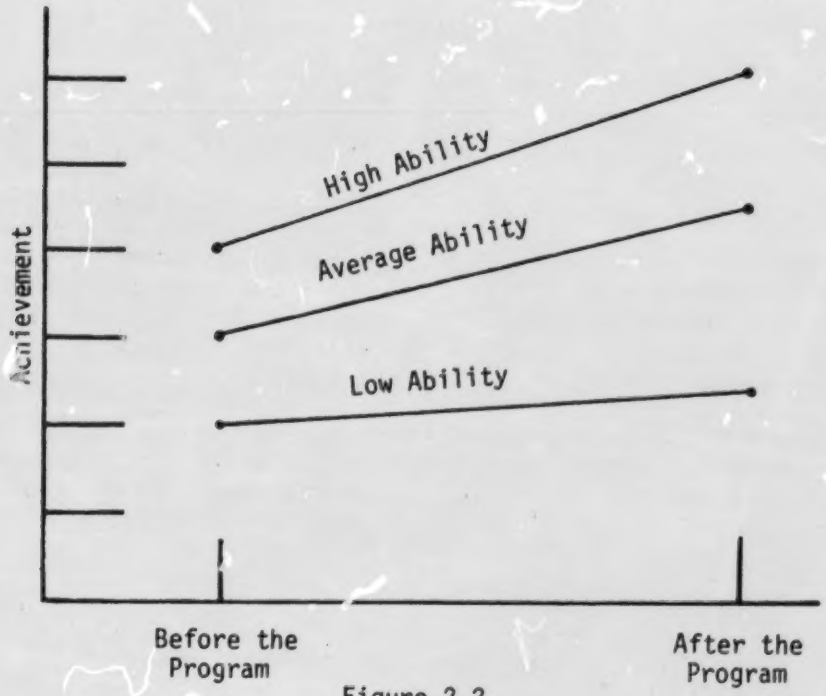


Figure 2.2

Trait-Treatment Interactions

An example of a trait-treatment interaction can be seen in Figure 2.2. Students of average and high ability experienced an improvement in achievement level during the program. The performance level of low ability students, however, remained virtually unchanged. Clearly the effectiveness of this particular program to bring about improvement in achievement interacts with (or is dependent upon) the ability level of the children involved.

When using test data for program evaluation purposes, it is always wise to examine it for evidence of trait-treatment interaction. Such information can often be invaluable when program revision and improvement is being undertaken.

Finally, a last point concerning the use of standardized tests for program planning and evaluation involves the issue of sampling. When a school or district wishes to use standardized test results to arrive at a reliable estimate of the performance of all students within a grade or program as a group, it is often not necessary to test all students in that group. Unless scores of individual students are wanted for some other purposes as well, various sampling procedures can be used to give a valid picture of group performance.

For example, a random sample of students or classes of students from

across the district may be asked to write a reading test to provide an estimate of how well the district as a whole is doing. Alternatively, all students are tested, but each student is given only a sub-set of all of the test questions (referred to as matrix sampling). Details of these and other issues concerning program evaluation can be found in the book edited by James Popham which is listed among the references at the back of this booklet.

2.2 What Guidelines Should be Applied when Selecting a Standardized Test?

The selection and use of a standardized test should rarely be an isolated event. Rather it is carried out in the context of a unified testing program that has identified policies and procedures. Some important issues related to the selection of standardized tests are thus determined by the policies of the testing program. These issues will be dealt with in detail in Chapter 6 of this booklet. This section will focus on technical issues to be kept in mind when selecting any standardized test.

Assuming that a decision has been made regarding the purpose to be served by the test, emphasis should then be placed on the instrument's technical qualifications. Essential to the determination of these technical aspects is access to a copy of the test, the user's manual, and any reviews of the test which may exist within Buros' Mental Measurement Yearbook or other publications listed in Appendix B.

There is perhaps no area where the dictum caveat emptor, "let the buyer beware," is more appropriate than in the selection of tests. Test publishing companies are in the business of selling tests, and it is the wise user who refers to any available outside reviews before making a selection. Remember, decisions affecting the future of both students and programs are going to be based in part of information provided by the test. As you select your test, you are in fact purchasing that information. Make sure you are getting precisely what you want.

There are three technical aspects in particular to be concerned with: validity, reliability, and adequacy of the norm-referencing.

2.2(a) Validity

Validity in a test is only partially contained within the test itself. The use to which the test results are put is the key factor in determining a test's validity. A scholastic aptitude test is valid, then, if it predicts school success well for your particular group. A diagnostic test is valid if it adequately samples relevant content you teach and produces guidelines for remediation you can implement. The test-makers, however, did not produce a test for the specific situation in which you will use it. In the most basic of meanings, then, it is the test-user who controls the extent to which a test will be valid, by selecting and using one that produces useful information.

The test-maker should produce evidence in the technical manual that the test produces valid results in the broader population. This process (called construct validity) assumes, at a minimum, that a test measuring "mathematical reasoning" produces results similar to other tests measuring the same construct. An additional, more stringent requirement would be that the

test of "mathematical reasoning" does not produce results similar to those obtained by a test of "clerical aptitude" because there is no theoretical or practical reason for assuming that these two attributes are related. A test exhibiting construct validity, then, should correlate well with tests measuring the same or related abilities and correlate poorly with tests measuring different or unrelated abilities. Often, the test-user will have to find objective sources other than the publisher's material for information about this important aspect of test validity. An essential source of such information for most commonly used tests is Buros' Mental Measurements Yearbook.

The test-user, however, can consult the test itself for another aspect of usefulness: content validity. Does the test measure important aspects of ability as defined by the school's program? If the curriculum in elementary reading is largely experiential, a standardized test measuring reading speed and vocabulary may not validly test the children's growth in this important subject. Before any test is selected for use, the people involved in implementing decisions based on the results must agree that the test is valid in content terms. It is a severe and often justifiable criticism to say of a test: It did not examine what we expected our children to learn.

For scholastic aptitude tests, however, the prime criterion of validity is not the content of the test but the degree to which any given test predicts future school success. Aptitude for school work is obviously an important factor in determining how successful a child will be in learning what the school asks him to learn. An aptitude test is useful (and valid) to the extent that the test results predict accurately the child's future performance on school-related tasks. This type of validity (predictive validity) has one important restriction: The use of scholastic aptitude tests alone to assign children to groups or grades can result in a self-fulfilling prophecy. Children who are expected to do well (or poorly) will tend to be seen as doing well (or poorly). Although aptitude for school work is an important factor in school achievement, others such as interest, perseverence, and attendance are also important, and the too heavy reliance on aptitude test results alone could seriously affect the educational futures of many children.

2.2(b) Reliability

Reliability in a test, like reliability in a watch, is the measure of the extent to which it measures accurately whatever it is that it measures: "achievement" or "aptitude" for a test; "time", for a watch. The score of any student on a test is an approximation of what he really knows. The more accurately the test measures this knowledge--the degree to which chance, selection of items and their consistency, and other non-test related factors are minimized--the more faith can be placed in the score awarded the individual or his group.

The key statistic to be examined in the test manual is the standard error of measurement. This number, derived from the reliability estimate and standard deviation of the test during its norming period, provides a "confidence band" for an individual or group score. If the standard error of measurement for a test (sometimes abbreviated to be SEM or s_e) is given as .4 grade equivalents, then a child's score of 5.4 grade equivalents should be seen as an estimate, bounded by $\pm$ one (or more for greater accuracy) SEM's. It is likely, on repeated testings, that the student would score differently than

5.4, and the SEM provides a "band" within which his scores are likely to cluster. This concept is discussed further in succeeding sections and in the glossary.

The usefulness of the SEM is that it illustrates the error that is always involved in measurement while at the same time providing a range of the possible breadth of that error. Generally speaking, the smaller the SEM is for a test, the more reliable the results. Teachers should use this concept as a check against over-interpreting the results; parents should have their children's scores reported to them as a width of scores rather than as a precise number.

2.2(c) Adequacy of Norm-Referencing

Both the validity and reliability information contained in a test are derived from the norm-referencing the test-makers performed in developing the test. The adequacy of the norm-referencing, then, is crucial to the test's success. If the samples of students selected by the publishers to serve as norming groups are not large, not representative, and/or not specific enough for the test's purposes, then the validity and reliability data may not be accurate nor useful for your purposes. Canadian test users often must make do with tests normed on American youth, a fact which may or may not hamper the usefulness of the tests. The development of local norms, along with local validity and reliability information, is one option to be considered when this situation arises. This process is described in more detail in Chapter 4.

CHAPTER 3

ADMINISTERING AND SCORING STANDARDIZED TESTS

When two or more tests manage to satisfy the technical criteria of validity, reliability, and adequacy of the norm-referencing, a choice among them may be made for administrative reasons. A simple rating of each test by those most involved in administering and translating the results into actions is one key means of insuring that the information generated by the test is used. In addition, there is the issue of organizing and implementing the testing, of how best to make the operation as efficient and painless as possible.

3.1 When Should Standardized Tests be Administered During the School Year?

Tests should be scheduled well in advance so that the disruption of regular classes will be minimized. It is important to avoid conflicts with exciting athletic events, field trips, parent days, etc. In general, tests should be given as near to major decision points as possible and thus the specific testing times will vary depending upon the level and extent of grades involved. Usually it is wise to administer tests at the time of the school year that was used in norming the tests.

There are two times during the school year when ability and achievement tests are generally administered. The first is in the Fall, a short time after school is open, and the second is in the Spring. The choice of timing depends upon the purposes for which the tests will be used. Generally, group readiness, ability, and interest measures should be given in the Fall period as the major goal of these tests is to provide information for instructional planning. Follow-up testing of individuals may be indicated from the analysis of group results, and the ability of teachers to provide remediation or enrichment may well depend upon the speed with which they can obtain this information.

Besides these instructional planning goals, Fall testing is useful where there is a high annual transfer of students resulting in inadequate achievement information from in-school sources. Where teachers may be concerned over the misuse of test data i.e. in evaluating their performance, Fall testing can alleviate anxiety and stop the "teaching for test" habit that hampers morale, test validity, and the curriculum.

Where the purpose of the testing is focused on program or course evaluation, Spring testing is likely to be more appropriate. For example, to evaluate the cumulative effects of ten years of schooling, survey achievement tests might be administered near the end of the junior secondary cycle. Similarly, testing of achievement in specific courses would of necessity be conducted after the course work is completed for the year.

Because such testing does occur at the end of the course or program, it is seldom of any direct use in helping students learn. Where standardized achievement tests are used in grading individual students, a clear misunderstanding of the tests' function exists. No commercially published standardized test is keyed closely enough to any one curriculum to permit such individualized interpretation. The same characteristic should also inhibit individuals from assessing a teacher's performance on the basis of this information. What these tests do measure are those general abilities that most schools and programs are ultimately interested in developing, and it is to the question of how the total school program is working to these ends that the results of standardized tests can provide useful information at the end of some defined period.

3.2 How Should Standardized Tests be Administered and Who Should do it?

For the actual testing program, some one person must be in a position to coordinate the operation. It may be possible, in some districts, to have outside personnel administer the tests to all the students. The benefits of this style are two: First, the training involved in ensuring that the testers are familiar with the procedures is easier with fewer people; and secondly, the use of fewer testers increases the probability that the tests will be administered to all pupils in the same way.

If teachers are used to administer the tests, there will be a saving of direct costs. Some pre-testing training, however, must be allowed for. As noted earlier, the relevance of the norms is dependent to a large degree on the understanding that all students complete the test under precisely the same conditions. For example, the time limits for sub-tests must be strictly adhered to; the environment must be free of outside disturbances; the importance of attempting to do well must be emphasized, and so on. If teachers are familiar with the goals and requirements of the testing through a training session, then the advantages involved in using teachers are maximized. By understanding what the purposes and requirements are, teachers are likely to be more interested in what the information shows that is generated by the results. A test administration guide (Appendix E) is provided that can be used to check the typical concerns involved in administering any standardized test.

3.3 How Should Standardized Tests be Scored and Who Should do it?

Tests should be scored accurately and as quickly as possible with a minimum expenditure of money, time and energy. Accomplishing these goals involves answering two questions: Should the test be scored by hand or machine? If scored by hand, who should do the scoring?

Standardized tests administered at the kindergarten-primary level are usually scored by hand. Some publishing companies, however, have developed test booklets for young children that can be machine scored. At the upper grade levels nearly all companies provide the option of having the test answered either in the booklet or on a separate answer sheet. When separate answer sheets are used, the booklet may be reused. Schools contemplating the use of separate answer sheets and/or machines scoring the test should

carefully consider the following factors.

(1) Many slow-learning pupils below the fifth grade will have difficulty with separate answer sheets unless extensive practice is provided before the tests are given.

(2) Each answer sheet must be scanned to detect marks that are too light, as well as cleaned of stray marks, before it goes to the scoring maching.

(3) Determine the time length between test administration and the return of test data to the school. A delay in returning data can be a serious shortcoming of machine scoring.

(4) Check the relevant costs of machines scoring booklets with consumable editions. In some cases the machines scoring booklets are considerably more expensive. They are printed on a better grade of paper to provide for their reuse and also, because of the lower value of sales, the net return to the publishers and to the authors would be greatly lowered if the prices were not higher.

(5) Consider the number of times the booklets would have to be used before any savings would be effected. This cost along with the burden of scanning and storing two or three forms of the test may greatly reduce the apparent economy of the machine scoring process.

(6) One feature that machine scoring services provide is the wide range of derived scores and accompanying statistical summaries. This service is only valuable, however, provided the school staff can use it to increase the quality of the education program. Unless this information is accompanied by some in-service activities, it may be money poorly spent.

Schools opting to score tests by hand must determine the most effective procedure to use. Many schools rely on classroom teachers to score the tests. Others employ trained clerks for the task--particularly above the primary grades. The trend appears to favour the latter for the following reasons:

i) Scoring objective answers is a clerical process, and trained clerks are likely to perform these tasks efficiently and reliably.

ii) With careful selection procedures clerical workers can be trained not only to score tests efficiently, but also to produce summary statistics and item analysis data.

iii) Teachers can spend their time interpreting, rather than generating, the information provided by the test.

Appendix E provides a complete, concise checklist for those involved in the actual supervision of testing activities.

CHAPTER 4

INTERPRETATION OF STANDARDIZED TEST SCORES

Raw scores on the vast majority of tests, including standardized tests, have little intrinsic meaning. One procedure for ascribing meaning to the raw scores is through referencing them to frequency distributions which describe the performance of specified groups. These distributions of scores and their raw score/converted score equivalents which permit interpretation of the relative standing of an individual's performance with reference to the group whose performance is described are called "norms". The test would be described as "norm-referenced" in that scores are interpreted by referencing them to the table of norms.

By contrast, criterion-referenced tests are usually interpreted in terms of an absolute standard. For example, a level of acceptable mastery of the content covered by the test may be defined in advance as 80%, and each student's performance is compared with this minimum pass level. Ideally, and particularly in the case of widely used published tests, one would expect that the mastery level would be based on empirical studies showing that the specified level has been determined to be necessary if students are to make satisfactory progress in the next unit of work. Unfortunately, in many of the currently available criterion-referenced tests, the empirical evidence is weak or lacking. A more detailed discussion of differences between norm- and criterion-referenced tests is provided in the manual "Construction and Use of Classroom Tests", as well as in other sources listed in the References at the back of this manual.

For our purposes, it is important to perceive the fact that these two types of tests perform quite different functions. Criterion-referenced achievement tests are best suited to individual student assessment, specifically for evaluating the degree to which students have learned the specific content they have been taught. Standardized achievement tests do not measure the specific content of instruction and can not, therefore, be used effectively as comprehensive diagnostic tools. Their purpose is to sample broader abilities, the wider goals of education, and to report how an individual or a group stands in relation to others. Detailed analysis of students' responses to individual criterion-referenced items should inform the teacher precisely what needs to be taught and to whom; responses to individual norm-referenced items are interesting primarily as they combine with other item scores to create a single score describing a broad area of achievement.

4.1 What are the Differences Between "Norms" and "Standards"?

It is also important to distinguish between the terms "norms" and "standards". Norms should not be invested with evaluative connotations: Being "below the norm" is often automatically equated with poor achievement, while "above the norm" is seen as commendable achievement. Norms are simply

descriptive tables, lacking evaluative connotations. Standards, however, do incorporate value judgements of the type permitting qualitative ratings of achievement, as "exceptional", "satisfactory", "failure", etc.

The usefulness of normative information depends upon the nature of the group or groups whose achievement is described by the norm tables. In general, particularly with achievement and ability tests, there may be several reference groups with which a student may be compared, depending upon the context and specific purpose of the comparison.

Standardized test manuals usually contain what are frequently called national norms. These are useful for comparing an individual score with those of a sample of students in various communities across the nation. Many of the standardized tests sold in Canada were normed in the United States which can decrease their appropriateness, particularly in the case of achievement tests when curriculum validity is of major importance. When selecting tests, the user should take particular care to examine the technical manual for information concerning the procedures used to norm the test.

Ideally, the technical manual should explain how students were chosen to take part in the norming process as well as describe in detail the proportion of pupils from such subgroups as the following:

- Boys and girls
- Geographical regions
- Rural and urban areas
- Socio-economic levels
- Racial groups
- Size and types of school

It is obvious, due to the large number of variables involved in developing norms, that scores on tests published by different companies are not directly comparable, even though the same converted scores (percentile ranks, grade level equivalents) are used. A standard score of 50 on a test published by Company X cannot necessarily be interpreted as being the same as a standard score of 50 published by Company Y. Some test companies, however, do use the same student sample for norming achievement and aptitude tests. This procedure is to be commended as it allows for direct comparisons of an individual's scores on both types of tests.

4.2 How can Local Norms be Established and Used?

Although national norms are of interest to administrators and curriculum personnel, they are of more limited use to the classroom teacher. The latter are concerned mainly with making decisions concerning selection and placement based upon how individuals compare with other students in their own school or district. For these purposes, local norms can be constructed in a relatively simple and straight-forward way. Since norms are basically frequency distributions, the procedure begins with (1) obtaining the appropriate individual scores either from a representative sample of students or from all students presently enrolled in that school, grade, or course; (2) completing a frequency distribution of those scores; (3) summarizing the data into one or more types of converted scores (discussed on following pages).

Local norms may be calculated for both teacher-made and for standardized tests that are used repeatedly.

One possible weakness of local norms is that the sample may be small thereby making the data unreliable. That is, with small groups of students, one runs the risk that the particular group enrolled in a course or grade this year may be atypical in some way from subsequent groups who may be given the test later. If the same standardized test is given yearly, it is possible to combine the scores of students over several years and thereby increase the confidence you can have in the stability of the norms.

4.3 What is Measurement Error?

A concept fundamental and of critical importance in the interpretation of any test score is "measurement error". Tests are measuring instruments (much as are meter sticks or weighing scales) designed to express in quantitative terms how much of some attribute the person being tested possesses. That attribute might be "ability to comprehend textual material", "knowledge of algebraic operations", "attitude toward school", or many others. It is assumed that the person tested has some true "amount" of that attribute. This hypothetical value, referred to as his true score, can be thought of as the average score on an infinite number of repeated tests that are uninfluenced by practice effects or memory of previous performance.

Tests, however, are fallible and the score obtained by the person at any one time can only approximate his true score. For example, error is introduced by virtue of the fact that the items included on a reading test only represent a small sample of all of the reading passages that might have been used and we cannot be certain how much these particular passages may have biased performance. As well, the student may have made minor errors in following directions on the test, or may have marked an incorrect response through inattention. He or she may have been anxious, tired, or subject to any number of other factors which influenced his performance.

All of these extraneous factors which interfere with the precise quantification of the amount of the attribute that the person possesses are referred to as measurement error. All test scores are made up of the true score plus measurement error. Therefore, in interpreting any test score, it is essential that it be viewed as an estimate only, recognizing that the same test given to the student on a different day, or a different test given the same day, may have resulted in a different score. For further information on this topic, the reader is referred to the discussion of Standard Error of Measurement in the Glossary at the back of this booklet and to the discussion on reliability in Chapter 2.

4.4 Why is the Normal Curve Used as the Basis for Most Standardized Test Scores?

One of the most common assumptions underlying testing in the schools (both standardized and teacher-developed) is that achievement and ability are somehow distributed in the population in the same way that height and length of forearm are distributed: A few individuals at each extreme with most of us

bunched tightly in the middle somewhere.

We may feel this way, in part, because test scores are often reported in this manner. Professional test-makers use the normal curve because so much is known about its characteristics, and these characteristics are useful in developing test statistics. The most relevant of these characteristics for our purposes are the following:

(1) In a normal distribution, the "average" or mean, the middle score or median, and the most common score, the mode, are all identical;

(2) In a normal distribution, one standard deviation* on each side of the mean will encompass the scores of approximately two-thirds of the individuals taking the test. (In fact, normal curve tables can clarify the position of any point on the curve in terms of percentiles but the broad standard deviation units are most frequently used.)

By using the normal curve with its known characteristics, test makers can easily make statements such as the following:

(1) There will be as many individuals scoring above the mean as below it;

(2) Less than 5% of the population will have scores greater or less than two standard deviation units away from the mean;

(3) A score of an individual that is one standard deviation above the mean can be quickly interpreted in a number of ways (see Figure 4.1): 84th percentile; T-score of 60; stanine 7; and so on. (The same procedure can be used for other scores at other points, and for scores derived from other tests using the normal curve as a basis.)

These uses have ensured that virtually all standardized tests produce scores that are "normal" yet virtually none of the test-makers assume that "achievement", for example, necessarily follows a normal pattern. Indeed, most educators agree that, as children increase in age, achievement becomes more variable as slower children fall farther behind their more talented classmates of the same age and grade.

The standardized test produces normal results, then, because the test-makers ensure this result by a judicious selection of items used to formulate the test. When many items are designed so that approximately one-half the students for whom the test is intended pass those items and one-half fail those items, a normal curve approximation will result. Another set of items with different characteristics would produce a far different distribution. Which distribution is the "correct" one is anyone's guess: We simply have no theoretical justification for preferring one over another. Until we do, the normal curve will be preferred for the reasons detailed above.

The fact that the test is engineered to produce variable results has several important ramifications for interpretations of the scores which result. Because the test-maker selects items which will discriminate among

*See Glossary for a technical description of this statistic.

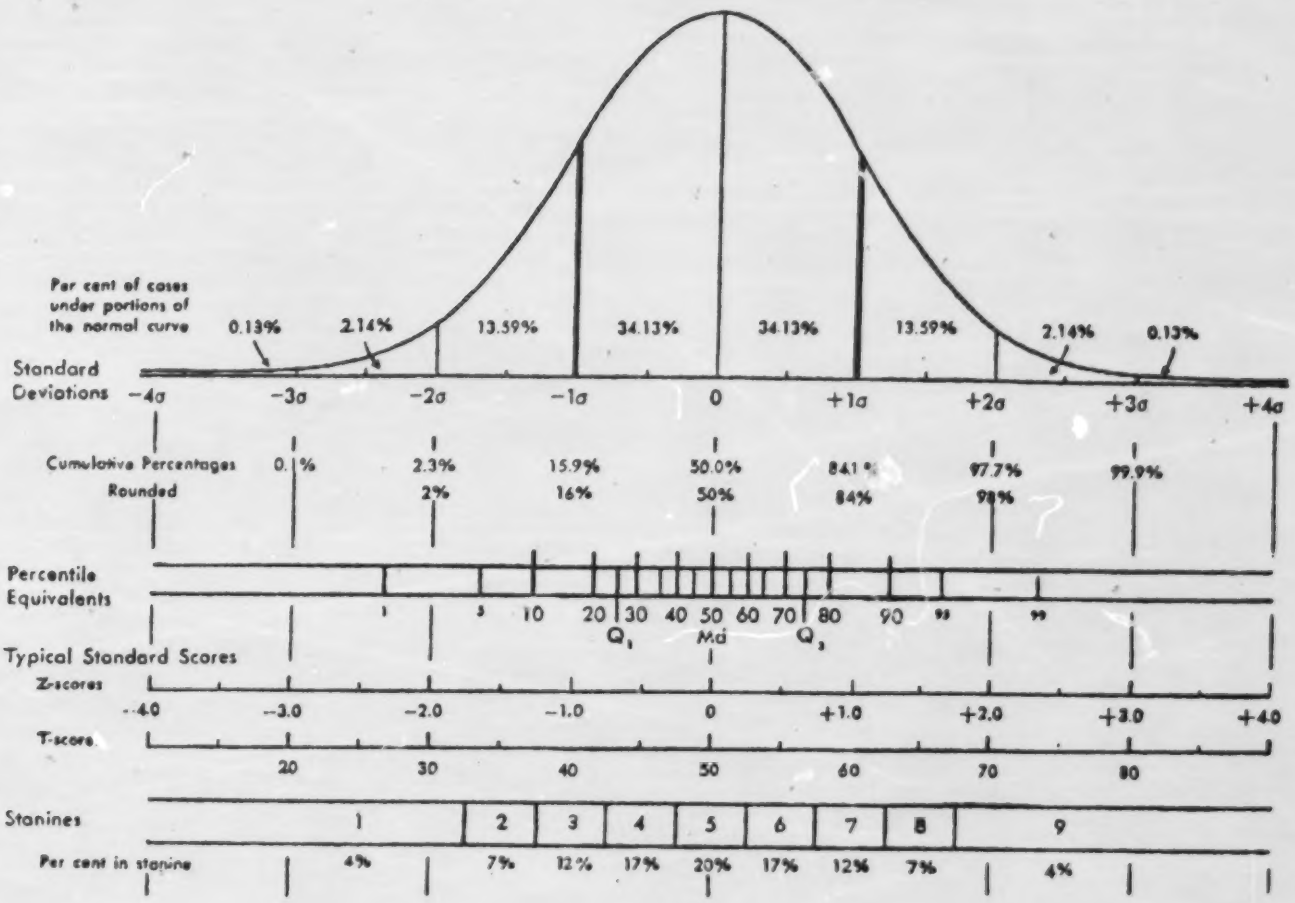


Figure 4.1

Relationship of Various Derived Scores in a Normal Distribution (Adapted from Harold G. Seashore, et al., Test Service Bulletin, No. 48. Psychological Corporation, January, 1955)

NOTE: In order to use the relationships shown in this figure, there must be evidence that the original raw score distribution is essentially a normal one.

learners, there is reason to expect that in any given sample of students, the variability in scores will be large. No educator can expect, therefore, that all students will score above the mean or above grade level. The tests are specifically designed to prevent this from occurring. In fact, if the norms for the test are perfectly matched to the sample, and as many students scored below their grade level as above it, the fair interpretation would be that this group of children is progressing in a way typical of that age and group.

An additional conclusion would be that variability in achievement is precisely what makes the standardized test useful. All norms tell how any student or class is progressing compared to other students or classes. The measure of distance away from the average, the standard deviation, is therefore the central piece of information necessary for test interpretation. It is the basis for all the derived or converted scores used in standardized tests. Ultimately, what a standardized test score says is: This is where Gary or Ian or Mr. Adams' class is in relation to the average.

The simplest way to perform this task would be to state, in standard deviation units (called z-scores), where Gary or Ian was in relation to the average. If Gary's score is +.75 and Ian's is -1.33, we could say that Gary was three-quarters of a standard deviation above the mean, and Ian was one and one-third standard deviations below the mean. Unfortunately, this simplicity begs the question of what that means.

It is here that the normal curve again becomes useful. Because tables are available that tell us what proportion or percentage of students must be included at each point along the distribution, we can translate any z-score into a percentile rank. For example, we know that on any test producing normal curve results, 16% of the students in the large norming sample will score at or less than -1.00 standard deviations (see Figure 4.1). Because the normal curve is symmetrical, as many students will receive scores above 1.00 standard deviation as students receiving scores below 1.00 standard deviation. If we choose to call scores below 1.00 "below average" (a standard, not a norm) and scores above 1.00 "above average", then the scores in between can constitute the average range, and this standard will encompass approximately two-thirds of all the students taking the test.

4.5 How Should Scores be Interpreted?

What we have done at this point is to convert the raw score of an individual into standard deviation units, and after accomplishing this, we have converted that score into a percentile rank or into any other form we wish based on the normal curve and norming group's characteristics. Various other converted scores are illustrated on Figure 4.1. A description of these, along with their advantages and disadvantages, follows.

4.5(a) Percentile Ranks

As described earlier, one of the most convenient methods for judging a student's score is to compare it to scores of other students within his own grade who took the test when norms were established. When a raw score is converted into a percentile rank, this rank states the percentage of scores in the norm distribution that the score under question equals or exceeds. For example, a percentile rank of 68 is better than or equal to 68 percent of the scores achieved by the norm group. When using percentile rank information, it is important to keep in mind that values between the 16th and 84th percentile are all within one standard deviation of the mean. Thus, while percentile ranks of 30 and 80 may appear to be very different, they are actually both within a relatively average range of performance.

Percentile ranks have one serious disadvantage. The size of the units of measurement along with the range of raw scores are not equal due to the

fact that typical achievement score distributions have a normal distribution: i.e. The raw scores increase in frequency at the centre of the distribution and then decrease in frequency at the lower and upper extremes. An apt analogy would be a meter stick that has centimeter markings which become increasingly wider apart as they approach the ends of the stick. This results in raw score differences of equal size leading to unequal differences in percentile ranks. Two students with percentile ranks of 50 and 60 may only differ from one another by 4 or 6 raw score points. Two other students with percentile ranks of 80 and 90 may differ from one another by 15 or 20 raw score points. Thus, although the percentile rank differences between these two pairs of students was 10, the two cases implied very different raw score differences.

The fact that equal percentile rank differences cannot be taken to represent equal differences in "amount" in an absolute sense, rules out their use for obtaining an average score for an individual based on a number of different tests. That is, because percentile ranks are in fact simply ranks, and are not based on an equal interval scale, they cannot be meaningfully added or subtracted. Other converted scores (such as standard scores) are more appropriate for combining test score data.

A caution must also be issued related to interpreting class or school averages in terms of the percentile ranks provided in the test manual. In order to convert a class or school average into a percentile rank, it is necessary that class or school norms be available in the manual. Some publishers only issue individual norms and, in such cases, it is incorrect to use the individual norms as a basis for interpreting group averages. This is due to the fact that the variability of individual achievement scores is much greater than school to school variability. The highest school average is considerably lower than the level of the ablest single student, and the lowest school average is considerably higher than the level of the least able student. Using individual percentile rank norms for interpreting group averages would therefore underestimate the relative achievement level of above average schools and overestimate the relative achievement of the below average schools.

4.5(b) Stanines

Stanine is an abbreviation of the term "standard nine", because these scores were derived by dividing the normal distribution into nine parts as shown in Figure 4.1. The fifth stanine represents the middle 20% of the scores (10% on each side of the mean) the 4th and 6th stanines include the next 17% of the distribution, and so on. The average range includes any scores falling approximately between the 3rd to the 7th stanine, although it will be noted from the figure that this average range does not include the entire 3rd and 7th stanines.

Grouping scores together and assigning a stanine to them results in a loss of information (i.e. percentile ranks of 40 and 50 both would be interpreted as a stanine of 5). However, considering the fact that all tests have measurement error associated with their scores, a stanine system can help convey the important concept that a score should be viewed as a band rather than as a precise point.

Stanine scores are assumed to be equal units of measurement. They can therefore be added across different achievement areas

and are also directly comparable across tests, provided that each test was normed on the same student group. For most reporting purposes, then, stanine scores are to be preferred. They are relatively simple to understand, and are less open to misinterpretation than other derived scores.

4.5(c) Standard Scores

Standard scores are somewhat more complex than those previously discussed. There are various types of standard scores, as well as different methods for transforming raw scores into standard scores. All methods, however, are based upon characteristics of the normal distribution. Details are provided in the Technical Glossary of this manual.

Technically, standard scores have a number of advantages over other types of converted scores. Particularly with standardized tests, the scores are considered to be equal units along the score scale. As such, several scores from different sub-tests can be combined into a composite measure. They can, like stanines, allow for direct comparison of scores on two or more tests but without the loss of score information that accompanies the grouping of scores in stanines. With their fixed mean (usually 50) and standard deviation (usually 10), standard scores provide one of the most precise placements of students along a common score scale.

T-scores, one type of standard score are shown on Figure 4.1. T-scores can range from 0 to 100, with scores from 40 to 60 falling between $\pm$ one S.D. These scores were developed from z-scores in order to eliminate the negative values associated with z-scores. Any student's z-score (a score, you will remember, expressed in standard deviation units) can be translated into a T-score by multiplying it by 10 and adding 50.

The intelligent quotient (I.Q.) is another example of a standard score. Its use is restricted to scholastic aptitude and other types of predictor tests. Originally it was computed as a ratio of a student's mental age divided by his chronological age, and then converted to a percentage. This procedure is no longer used. Instead, raw scores are changed into standard scores at each age level, so as to have a mean of 100 and a standard deviation of 15 or in some cases 16.

When standard scores (including I.Q. scores) are used with normal distributions (in most cases standardized tests, when administered to a large cross section of students, are designed to display a normal distribution) there is a direct and definite relationship between the I.Q. score and the percentile rank, i.e., an I.Q. of 100 equals a percentile rank of 50. Thus when one knows the relationship between I.Q. scores and percentile rank scores, the interpretation of I.Q. scores becomes more meaningful. Similarly an I.Q. of 125 signifies the same status, relative to the student's own group, whether obtained in grade 4 or 7.

4.5(d) Grade Equivalent Scores

One of the most popular types of converted scores used with Achievement Tests is the Grade Equivalent Score (G.E.S.). It is defined as the grade level for which a given raw score is the real or estimated average. Due to the fact that schools operate only 10 months during the year, G.E. scores are expressed in 10 one-month intervals; that is, norms for grade VI would

include G.E. scores ranging from 6.0 to 6.9 in one-month intervals from the first to the last month of the school-year. If Ann obtained a raw score which converted to a G.E. score of 6.5 this could be interpreted as "performing as well as the average student who has completed 5 months of the 6th grade".

Despite their popularity and apparent simplicity in interpretation, Grade Equivalent Scores present a number of shortcomings of which the following three are noted. (1) Standardized tests, during the norming process, are usually given at the same time of the year to not more than three different grade levels. Consider a particular test designed for grades 4-6. Let us assume the test is administered and normed to a representative group of students in October of a given year. The average raw scores for grades IV to VI are calculated to be respectively 25, 35, and 40. A norm table using the actual observed score values is produced below.

Score	Grade Equivalent
25	4.1
35	5.1
40	6.1

Assuming that the maximum score on the test is 50, a problem arises as to how to equate the other possible raw scores (0-24, 26-34, 36-39, 41-50) into Grade Equivalent scores. The problem is solved by a process of interpolating between the various mean scores and extrapolating below score 25 and above 40. This process is done graphically and involves a considerable amount of arbitrary judgement. Note that the extrapolated scores will all be below 4.1 or above 6.1 even though no students at these grade levels actually wrote the test. It must be constantly kept in mind that Grade Equivalent scores beyond the actual tested grade limits are meaningless. Raw scores well into the extrapolated regions imply that the test was too easy or too difficult for the students' scoring at these levels, and that for these students, tests of the next higher or lower level should be used.

(2) When raw scores for the norm group are converted into Grade Equivalent scores, the mean or average raw score for the group is labeled as the grade equivalent. In the example above, the mean raw score for grade 4 in October was 25. Hence this score is labeled 4.1 in Grade Equivalent terms. What must be recognized is that if 25 is the mean, then approximately one half of the students' raw scores were below 25 and the other half were above. Thus, the norms are set in such a way that approximately one half of the children would have grade equivalent scores which are lower than the grade level at which they are now actually assigned in school. Many users of Equivalent scores treat these as absolutes and fail to recognize the breadth of an average range. For example, in one widely used achievement test battery, the ranges between -1.00 and +1.00 standard deviations are as follows:

Subtest	Grade Equivalent Range $\pm$1.00 S.D.
Reading	2.3 - 5.3
Language	2.5 - 5.7
Mathematics	3.0 - 4.8

A grade equivalent score of 2.4 and 5.2 in Reading on this test are therefore both likely to occur in a typical class of students.

(3) Growth in different school subjects proceeds at different rates depending upon the in-school emphasis and out-of-school learning. The level of achievement in elementary school arithmetic is mainly a function of the amount of in-school time spent on the subject. However the level of achievement in reading can be greatly influenced by the student's out-of-school experiences and leisure-time reading activities. The different growth rates across subject areas affect the variability of the Grade Equivalent scores within a grade as well as across grades. The differences are depicted in the table below which report the standard deviation of Grade Equivalent scores for various subjects on a popular standardized test.

Standard Deviations in
Grade Equivalents for Various Subjects

Test	Grade Level		
	Grade 3	Grade 7	Grade 8
Reading	1.0	1.4	1.9
Spelling	1.2	1.6	2.2
Arithmetic	.8	1.1	1.7

It can be seen that the standard deviations increase with each successive grade and that the variability in reading and spelling is consistently larger than for arithmetic. The reasons for these differences were suggested earlier. They may also be due to the fact that arithmetic is generally taught sequentially and without encouraging diversity. Reading and spelling on the other hand display greater individual differences after students have learned the basic language arts skills.

If a student in Grade V (5.0) was at the 98th percentile on all three tests, (two standard deviations above the mean) his G.E. scores would range from 7.2 for arithmetic to 8.2 for spelling. It is obvious that comparing G.E. scores across subjects is often misleading and sometimes meaningless unless one knows their means and standard deviations. This is particularly true at the extreme ends of the norm distributions. In a popular standardized test, for example, a score of 29 correct in study skills gives a G.E. of 7.6. If a student obtains one more item correct the G.E. changes to 9.2. One cannot imagine that the learning that takes place in one year and six months can be actually represented by one test item. However a naive use of G.E. scores would lead to such an interpretation!

In some test relationships a student's G.E. score on one subtest may be higher than the other subtest even though his actual norm-referenced rank is just the reverse. For example, a student in grade V who obtains a G.E. of 6.3 in arithmetic, and a G.E. of 6.5 in reading may actually have a higher percentile rank in arithmetic than in reading. Beware of attempting to compare the noncomparable.

One of the most difficult tasks for a teacher is to correctly interpret a Grade Equivalent score to a student or parent. No parent wants to be told his son in the 8th grade is functioning at the 6th grade level. Nor is it easy to explain to a parent why their bright 4th grader who obtained a G.E. of 7.0 should not be moved into the 7th grade. The advanced 4th grade student does not know the same amount or the same kinds of things as a typical 7th grader. The high G.E. does not mean that the 4th grader has mastered the subject matter of the 7th grade. His score is the result of superior mastery of the 4th grade test items and perhaps correctly answering some of the easier items designed for higher grades. Due to the danger of misinterpreting Grade Equivalent scores, the teacher should use percentile ranks within-grade or stanine scores as the preferred methods of reporting standardized achievement test results to parents.

A particular advantage of Grade Equivalent scores is that they help teachers appreciate the great magnitude of individual differences that exist within a single grade level. If one plots on a graph the distribution of Grade Equivalent scores for grades 3-8 one can vividly picture the degree of achievement overlap between grades and the increased variability of achievement with each successive grade level.

Because of the limitations suggested above, Grade Equivalent scores should not be used for interpreting individual student's scores. However, conversion of a class or total grade mean score into a Grade Equivalent score will help determine whether the test was designed at an appropriate level of difficulty for the group. For example, if a grade 5 class is administered a standardized test in early October and the mean score equates to G.E.= 5.1 the teacher can be confident, provided the class is average in school ability, that the test will provide data for making reliable norm-referenced measurements. It must be noted in a case like this, that one half of the class scored below their grade placement (5.1). Knowing the way a Grade Equivalent score is calculated, this fact should not cause alarm amongst the teaching staff.

CHAPTER 5

REPORTING RESULTS OF STANDARDIZED TESTS

The purpose of using standardized tests is to provide information; specifically, information that can be used to decide courses of action that will improve the educational enterprise. At least seven groups can be considered prime users of this information, although not all groups need reports of every test administration. These groups are: pupils, teachers, parents, curriculum workers, administrators, school board members, and the general public. Not all these groups are necessarily familiar with the characteristics of standardized tests that have filled most of this resource book thus far. Reporting results of these tests to such diverse groups, therefore, requires much care if misinterpretations are to be avoided.

5.1 How can Misinterpretation be Avoided?

The results of a standardized test program are of two types; statistical and educational. Most of the interested parties to the results will be concerned with the latter; most of the misinterpretations will occur with the former. Although it is clear that the statistical results should be available for inspection, most reports should rely most heavily on the educational implications of the statistical data.

5.2 What are Some Alternative Reporting Formats?

Different groups require different reports if they are to assimilate the information wisely. Differing formats will also protect against the misuse of information. It may be wise here to re-cap certain principles that have formed the basis of this resource book.

First of all, tests are used to accomplish some purpose: to aid in diagnosis and remediation, to facilitate program planning and evaluation, to help place students within programs, or to aid in vocational planning. The test results, in combination with other information, should be reported in the light of that decision process which was the purpose of the evaluation.

Secondly, the information is fallible. There is no need to produce reports which allow for the invidious comparison of school to school, for example, because the scores on a general achievement battery are not reliable guides to the causes of variations in school achievement. Similarly, teachers should have access to the scores of the children in their classes, but do not need to know the data for other classes, in their own or other schools. What constructive purpose could be served by such a report?

In addition, administrators and curriculum workers are most frequently interested in cross-district achievement. If such information is the sole purpose of the testing, then multiple-matrix sampling procedures should have been used that actually prevent the misapplication of the test data for such purposes as informal teacher evaluation. Reports of the testing then can be made on a district basis without the possibility of class and school breakdowns which may lead to these undesirable outcomes.

Board members and the general public are also most interested in general rather than specific concerns. By concentrating on what the test results mean in terms of the schools' programs, the public's right to be informed is protected without encouraging arbitrary and fallacious misuse of information. Educational officials should not miss the opportunity to report how the educational system is functioning; it provides the basis for the continuing support of the schools by the community.

In summary, the format of reports may be different for different members of interested groups. For those groups interested in the broad findings (curriculum workers, administrators, board members, and the general public), the format should focus on the following:

(1) the purpose to be addressed by the evaluation;

(2) the major findings of the testing;

(3) how these findings correlated with other information relevant to the purpose;

(4) what actions flow from the total array of information.

5.3 How Should Information Best be Passed on to People?

Most standardized test manuals contain detailed sections on how to use and interpret test results. Usually special forms are available to facilitate proper interpretation to a student and his parents. These special forms may be adapted by teachers and others for use in local situations that are unique. When holding conferences with students and parents, teachers should use the test results only to supplement other available information such as what is known about the student's aptitude and about his educational and vocational goals. Whenever possible, truthful, accurate and factual information should be presented in a way that will lead to positive and construction action. When low test results have to be reported, the discussion should center on steps for possible remediation and improvement.

A common format for reporting standardized test results is a combination of narrative and graphic presentation. Profiles of student achievement on a standardized battery are generally reported as bands representing the actual score of the student (in percentiles or stanines) bounded by one SEM above the score and one SEM below the score. An example is given on the following page.

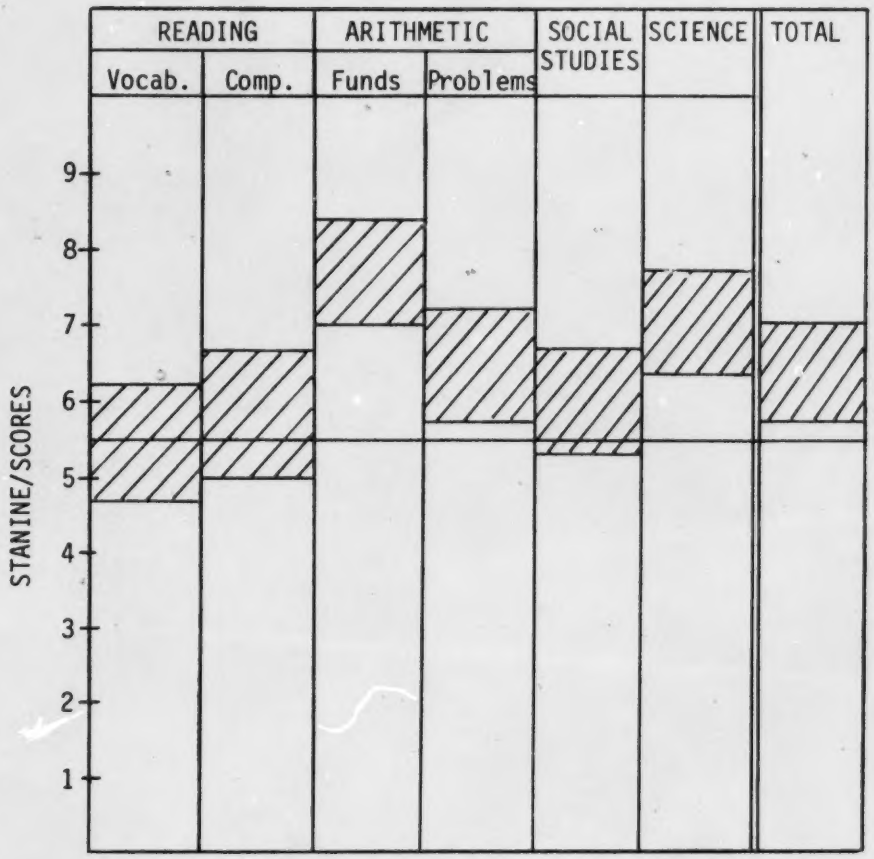


Figure 5.1

Profile of a Student's Achievement on a Standardized Achievement Battery

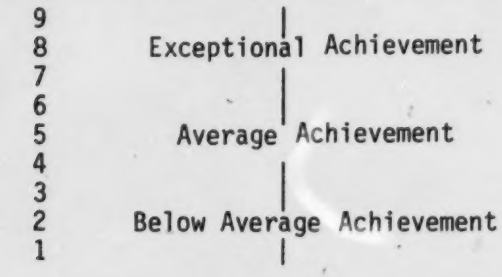
The information for parents should include a description of what stanine scores mean as well as a narrative relating their child's achievement.

Example: The attached graph represents Eric's achievement on the QED achievement tests which are administered annually to the Grade 8 students in our district.

The tests measure basic skills in the subject areas shown. The stanine scores represent your child's achievement compared to other Canadian children in his age group.

Because these tests are subject to error, each score is shown as a band representing the area your child's true achievement is likely to be found.

The stanine scores can be interpreted as follows:



According to the information on these tests and other information we have collected, Eric is demonstrating average ability for a child his age in Reading Vocabulary and Comprehension. His achievement in Arithmetic Fundamentals and in Science is above average reflecting his demonstrated interest in these areas. Eric's achievement in Arithmetic Problems is slightly above average, but might improve with more effort given Eric's achievement in Arithmetic generally. Overall, Eric is achieving satisfactorily when compared to other children his age.

Such a report might serve as a basis for a parent-teacher interview or for a "special" written report, as well as documentary material for the student's file.

CHAPTER 6

ESTABLISHING A STANDARDIZED TESTING PROGRAM

There is no single clearly defined testing program which can be used effectively in all schools. The rate and level of instruction, the unique goals of individual schools, differences in curriculum, size of schools, the characteristics of the student body, and other variables all make a single testing program unworkable. Each school district--in fact, each school--must develop the procedures that will best serve the various groups who make use of test data.

6.1 What are the Essential and Optional Components of a Testing Program?

Apart from this diversity, however, there are a number of elements common to most good testing programs. At the elementary school level a considerable amount of pupil evaluation is accomplished through use of techniques such as observation and checklists as well as anecdotal records and teacher made tests. These techniques, because they supply most of the information on individual student progress, form the basis of a good testing program. The addition of standardized achievement, school ability, and diagnostic tests to the program should depend on each pupil's growth and development. Only personnel involved in the schools' educational program can decide what and when additional test data are needed. As a basis for discussing a minimum standardized testing program at the school level the following sequence is suggested. However, no standardized tests should be administered, unless a detailed statement is prepared on how the results are to be used by various personnel to improve the instructional activities of the school.

Standardized Testing Program: (An Example)

Grade Level	Type of Test	Major Purpose(s)
K or 1	Readiness	- aid in organizing sub-groups for reading instruction - baseline for measuring individual progress
2 through 10	Achievement Battery	- program evaluation and planning - identifying students needing further diagnosis and remediation - additional data for teacher evaluation of student achievement
2, 3	Aptitude	- selection of atypical students for special guidance and placement

Grade Level	Type of Test	Major Purpose(s)
5, 6	Aptitude	- check on previous aptitude test results
9, 10	Aptitude - General - Specific	- selection for horizontal and vertical placement - educational counselling
10 through 12	Interest Inventories	- vocational guidance
11, 12	Subject Achievement (e.g. Algebra 12, Chem. 11/12)	- program evaluation and planning - normative student achievement

NOTE: In addition to these tests, additional information more closely allied to the specific instruction would be carried out through teacher-prepared tests, indirect assessment through observation, provincial assessments and so on.

The nature of the curriculum changes radically from the elementary to the secondary school. In the elementary grades the curriculum is quite similar for nearly all students, but at the secondary school the number of courses and program options greatly increase. The standard curriculum for all secondary school students has virtually disappeared. As such, standardized tests, particularly in the content areas, are very difficult to develop beyond grade 8. This is the major reason why Grade-Equivalent scores lack meaning in the upper grades of the secondary school. For these various reasons teacher-made tests based upon the local curriculum will continue to be the most common type of measurement technique used in the secondary schools. However, as students progress through the secondary school they must be aided in making decisions about future educational and vocational plans. In particular, school guidance personnel must not neglect the diverse needs of both the university bound and non-university bound students. Tests in the areas of special aptitudes and general interest may prove helpful in aiding students to plan their future goals.

6.2 What Steps are Involved in Planning a School Testing Program?

The most significant point to bear in mind is that an effective testing program concerns the entire staff of the school and demands the enthusiastic support of everyone involved. Tests can never be more than instruments to gather information needed to carry out other school functions. They are means to this end and can never be considered ends in themselves. The extent to which a testing program fulfills its purpose is therefore dependent upon the well organized administration of the program and the valid uses made of the test results.

The detailed organization of a testing program is dependent upon a number of factors such as the past experience of the teachers in testing programs, the size and the level of the school or schools involved, whether or not specially trained personnel are available etc. As such it is impossible to list detailed suggestions for each school. However the following general principles are

offered as guides toward developing an effective program.

(1) Testing Programs should be developed cooperatively by a standing committee of teachers, administrators and specialists in measurement and evaluation who are familiar with the curriculum.

(2) The committee should begin its work by establishing objectives for the testing program that, when fulfilled, will provide all interested groups with information to help make the learning environment better for pupils.

(3) Consideration should be given to the possible benefits of including pupils and parents on the school or district testing committee.

(4) To achieve continuity and integration of the program, the committee should represent teachers at all grade levels as well as supervisors in special subjects and areas such as reading, arithmetic and counselling.

(5) Although it is extremely important that many groups be represented in planning so that a variety of view points is obtained, in the long run, competent planning is more important than cooperative planning. The final detailed decisions about selecting particular tests, therefore should be the responsibility of the professional staff. As such, consideration should be given to the benefits of appointing a specialist to manage the testing program and other testing activities such as in-service education of school personnel in administering, interpreting and using test results.

A testing committee will have a number of responsibilities for planning and directing the testing program. First they must decide how and by whom the test results are to be used. Next they must decide what tests will be given (locally-developed, provincially-sponsored as well as standardized tests should be considered) in what grades and at what time of the year. Lastly they must recommend policies for scoring the tests and interpreting and reporting the results. These responsibilities will be discussed in the following sections.

6.3 How Should Tests be Selected?

Assuming that a decision has been made as to the types of tests that will supply information for more effective decision making, the question arises as to which specific tests should be used at each grade level. Considerations of importance are the validity of the tests (according to the local curriculum) and their reliability (preferably demonstrated in terms of a local pilot program). Other factors that should be considered by the committee include the cost, time limits, ease of scoring, effectiveness of norms and convenience of accessory materials for facilitating the use of the test results. It will be necessary for the committee to utilize considerable resource material during the test selection process. Many books on educational measurement contain chapters offering specific suggestions, including sample test evaluation forms. Buros' Mental Measurements Yearbooks are essential. Publishers' test catalogues give brief descriptions of their tests and specimen sets should be ordered for careful examination. This part of the committee's job entails a considerable amount of detailed study.

In general the following criteria should be considered when deciding what specific tests to use.

(1) All tests chosen should match well the objectives and curricular emphasis of the school. Each test chosen should contribute new and different information deemed helpful for identifying and solving practical educational problems.

(2) Whenever possible the same achievement tests series should be used throughout the grade levels to ensure the comparability of norms and sub-tests. This procedure permits comparison of student performance one year with their performance in successive years. Comparing scores of tests from one publisher with scores on tests of a second publisher is almost useless due to the differences in content organization, standardization samples and methods of expressing scores. When it is deemed necessary to change tests the procedure is to phase out the original test one grade level per year beginning at the lowest grade and introduce the new test at the grade level where the old test has been phased out. This allows the use of the same test for any given age group straight through the length of the test series, but at the same time allows schools to change publishers if they so desire.

(3) Whenever achievement and ability tests are administered to the same students it is strongly recommended that tests be chosen which were normed on the same standardization sample. This is particularly important when one wishes to compare the test results and determine the extent to which pupils are achieving up to their expected levels of performance.

Appendix F provides a convenient checklist for selection bodies to use in their deliberations.

6.4 What Procedures Should be Followed to Evaluate a Testing Program?

The preceeding sections have described and explained some of the problems associated with the development, selection and use of tests in schools. Foresight in planning to meet these problems will do much to ensure the success of any testing program. The following checklist will help focus attention on these problems.

Some of the items in the checklist may not be germane to specific situations in given schools or districts. If this is the case they should be noted and quickly passed over. A discussion of those items which are applicable will focus the attention of the appropriate personnel on the solution of any existing problems. If this occurs, the check list will have served its purpose.

Checklist on the Organization and Administration of the Testing Program

This checklist is divided into seven parts each labelled with a capital letter. The numbering of the questions is continuous throughout the entire checklist.

	YES	NO
A. Coordination, Leadership and Philosophy		
1. Is there an active testing committee representative of some or all of the staff of the school/district?	___	___
2. Is there a written statement of the duties of this committee?	___	___
3. Do members of this committee have course work in the area of tests and measurements?	___	___
4. Is there a system-wide testing director?	___	___
5. Is there a written statement of the purposes of the school/district's testing program?	___	___
6. Are these testing program purposes in harmony with the stated philosophy and objectives of the whole school program and with school and district policies on assessment?	___	___
7. Are school board members, parents and other citizens well-informed as to the extent and purposes of the testing program?	___	___
B. In-service Training Programs Related to Testing		
8. Are in-service training opportunities provided for those who administer, score, and interpret standardized tests?	___	___
9. Are in-service opportunities provided teachers for improvement of teacher-made classroom tests?	___	___
10. Does the school staff library have references on areas covering teacher-made classroom tests and uses and interpretation of standardized tests?	___	___
11. Are specialists in the area of tests and measurements available for assistance with problems as they arise?	___	___
C. Selection of Tests		
12. Are all tests selected by recognized professional and technical standards?	___	___
13. Do those selecting tests have specimen sets available for study and such critical reviews of tests as appear in Buros' Mental Measurements Yearbooks?	___	___
14. Whenever practical, are tests field-tested on small samples of students before broad application?	___	___
15. Are all tests based on objectives that are appropriate for the ability level and instructional program of the students involved?	___	___

	YES	NO
16. Is there a periodic evaluation of each test in the program to determine whether it should continue to be used?	___	___
D. Testing Facilities		
17. Are rooms where testing takes place satisfactory in terms of		
a. freedom from outside distractions?	___	___
b. adequate lighting?	___	___
c. adequate ventilation?	___	___
d. work space for each student?	___	___
18. Are there adequate, carefully supervised facilities for storage and control of testing materials?	___	___
19. Are all testing supplies subject to continuing inventory and checked as to usefulness?	___	___
E. Administration and Scoring of Tests		
20. Are teachers involved in determining when tests will be administered and for what purpose?	___	___
21. Is the person administering the test always prepared for the task?	___	___
22. Are the purposes and importance of each test made clear to students before the test is given?	___	___
23. Are persons present to assist with the administration of tests to groups when desirable?	___	___
24. Is the timing of the testing disruptive of the other activities of the school?	___	___
25. Do teachers spend an inordinate amount of time scoring, tallying, graphing and filing test results?	___	___
26. Are test scorers always carefully instructed for their tasks?	___	___
27. Are all student answer sheets carefully checked for possible errors in scoring?	___	___
28. Are services of a trained psychologist or psychometrist available for testing individual students?	___	___
F. Facilities and Means for Use of Test Results		
29. Are test results filed in places easily available to the persons who should use them?	___	___
30. Are definite means employed to encourage wider and more thorough use of test results?	___	___

	YES	NO
31. Are there rooms, private and quiet, in which teachers and others can confer with pupils?	___	___
32. Are cumulative test records carefully reviewed periodically, so as to be sure that test data and other material are properly organized and as easily useable as possible?	___	___
33. Is a test handbook describing the testing program and covering such areas as uses and interpretation of test information readily available for teacher reference?	___	___
34. Are teachers encouraged to make use of the item analysis technique for studying difficulties of individuals and their classes as a whole?	___	___

G. General

35. What are the strong points in the testing program?

36. What are the weak points in the testing program?

37. What improvements can be made in the testing program during the coming year?

REFERENCES AND SUGGESTIONS FOR FURTHER READING

Ahmann, J.S. and Glock, M.D. Evaluating Pupil Growth, 5th Edition. Allyn and Bacon, Inc., 1975.

Aitken, L.R. Psychological Testing and Assessment, 2nd Edition. Boston, Allyn and Bacon, Inc., 1976.

American Psychological Association. Standards for Educational and Psychological Tests and Manuals. 1200 Seventeenth St. N.W., Washington, D.C. 20036. 1974.

Brown, F.G. Principles of Educational and Psychological Testing, 2nd Edition. Toronto. Holt, Rinehart and Winston, 1976.

Mehrens, W.A. and Lehman, I.J. Standardized Tests in Education, 2nd Edition. Toronto. Holt, Rinehart and Winston, 1975.

Popham, W.J. (Ed.) Evaluation in Education: Current Applications. Berkeley, California. McCutchan Publishing Co., 1974.

Sax, G. The Use of Standardized Tests in Evaluation. In W.J. Popham (Ed.). Evaluation in Education: Current Applications. Berkeley, California. McCutchan Publishing Co., 1974.

Thorndike, R.L. and Hagen, E.P. Measurement and Evaluation in Psychology and Education, 4th Edition. John Wiley and Sons, 1977.

Ward, A.W., Backman, M.E., Hall, B.W., and Magure, J.L. (Eds.) Guide for School Testing Programs. National Council of Measurement in Education. Michigan State University, East Lansing, Michigan.

Wardrop, J.L. Standardized Testing in the Schools: Uses and Roles. Monterey, California. Brooks/Cole Publishing Co., 1976.

APPENDIX A

A GLOSSARY OF MEASUREMENT TERMS

The terms defined are the more common or basic ones such as occur in test manuals and educational journals. In the definitions, certain technicalities and niceties of usage have been sacrificed for the sake of brevity and, it is hoped, clarity.

academic aptitude The combination of native and acquired abilities that are needed for school learning; likelihood of success in mastering academic work, as estimated from measures of the necessary abilities. (Also called scholastic aptitude, school learning ability, academic potential).

achievement test A test that measures the extent to which a person has "achieved" something, acquired certain information, or mastered certain skills - usually as a result of planned instruction or training.

aptitude A combination of abilities and other characteristics, whether native or acquired, that are indicative of an individual's ability to learn or to develop proficiency in some particular area if appropriate education or training is provided. Aptitude tests include those of general academic ability (commonly called mental ability or intelligence tests); those of special abilities, such as verbal, numerical, mechanical, or musical; tests assessing "readiness" for learning; and prognostic tests, which measure both ability and previous learning, and are used to predict future performance - usually in a specific field, such as foreign language, shorthand, or nursing.

Some would define "aptitude" in a more comprehensive sense. Thus, "musical aptitude" would refer to the combination not only of physical and mental characteristics but also of motivational factors, interest, and conceivably other characteristics, which are conducive to acquiring proficiency in the musical field.

battery A group of several tests standardized on the same sample population so that results on the several tests are comparable. (Sometimes loosely applied to any group of tests administered together, even though not standardized on the same subjects.) The most common test batteries are those of school achievement, which include subtests in the separate learning areas.

bivariate chart (bivariate distribution) A diagram in which a tally mark is made to show the scores of one individual on two variables. The intersection of lines determined by the horizontal and vertical scales from cells in which the tallies are placed. Such a plot provides frequencies for the two distributions, and portrays the relation between the two variables as a basis for computation of the product-moment correlation coefficient.

ceiling The upper limit of ability that can be measured by a test. When an individual makes a score which is at or near the highest possible score, it is said that the test has too low a "ceiling" for him; he should be given a higher level of the test.

---

[1] *Reproduced in part from A Glossary of Measurement Terms. (Test Service Notebook, No. 13). Distributed by The Psychological Corporation.*

coefficient of correlation A measure of the degree of relationship or "going-togetherness" between two sets of measures for the same group of individuals. The correlation coefficient most frequently used in test development and educational research is that known as the Pearson or product-moment r. Unless otherwise specified, "correlation" usually refers to this coefficient, but rank, biserial, tetrachoric, and other methods are used in special situations. Correlation coefficients range from .00, denoting a complete absence of relationship, to +1.00, and to -1.00, indicating perfect positive or perfect negative correspondence, respectively. See Correlation.

composite score A score which combines several scores, usually by addition; often different weights are applied to the contributing scores to increase or decrease their importance in the composite. Most commonly, such scores are used for predictive purposes and the several weights are derived through multiple regression procedures.

concurrent validity See validity (2).

construct validity See validity (3).

content validity See validity (1).

correlation Relationship or "going-togetherness" between two sets of scores or measures; tendency of one score to vary concomitantly with the other, as the tendency of students of high IQ to be above average in reading ability. The existence of a strong relationship - i.e., a high correlation - between two variables does not necessarily indicate that one has any causal influence on the other. See coefficient of correlation.

criterion A standard by which a test may be judged or evaluated; a set of scores, ratings, etc., that a test is designed to measure, to predict, or to correlate with. See validity.

Criterion-referenced (content-referenced) test Terms often used to describe tests designed to provide information on the specific knowledge or skills possessed by a student. Such tests usually cover relatively small units of content and are closely related to instruction. Their scores have meaning in terms of what the student knows or can do, rather than in their relation to the scores made by some external reference group.

criterion-related validity See validity (2).

Curricular validity See validity (2).

deviation The amount by which a score differs from some reference value, such as the mean, the norm, or the score on some other test.

diagnostic test A test used to "diagnose" or analyze; that is, to locate an individual's specific areas of weakness or strength, to determine the nature of his weaknesses or deficiencies, and wherever possible, to suggest their cause. Such a test yields measures of the components or subparts of some larger body of information or skill. Diagnostic achievement tests are most commonly prepared for the skill subjects.

distribution (frequency distribution) A tabulation of the scores (or other

attributes) of a group of individuals to show the number (frequency) of each score, or of those within the range of each interval.

error of measurement See standard error of measurement.

expectancy table ("expected" achievement) A term with two common usages, related but with some difference:

(1) A table or other device for showing the relation between scores on a predictive test and some related outcome. The outcome, or criterion status, for individuals at each level of predictive score may be expressed as (a) an average on the outcome variable, (b) the percent of cases at successive levels, or (c) the probability of reaching given performance levels. Such tables are commonly used in making predictions of educational or job success.

(2) A table or chart providing for an interpretation of a student's obtained score on an achievement test with the score which would be "expected" for those at his grade level and with his level of scholastic aptitude. Such "expectancies" are based upon actual data from administration of the specified achievement and scholastic aptitude tests to the same student population. The term "anticipated" is also used to denote achievement as differentiated by level of "intellecutal status."

face validity See validity (1).

frequency distribution See distribution.

group test A test that may be administered to a number of individuals at the same time by one examiner.

individual test A test that can be administered to only one person at a time, because of the nature of the test and/or the maturity level of the examinees.

internal consistency Degree of relationship among the items of a test; consistency in content sampling. See split-half reliability.

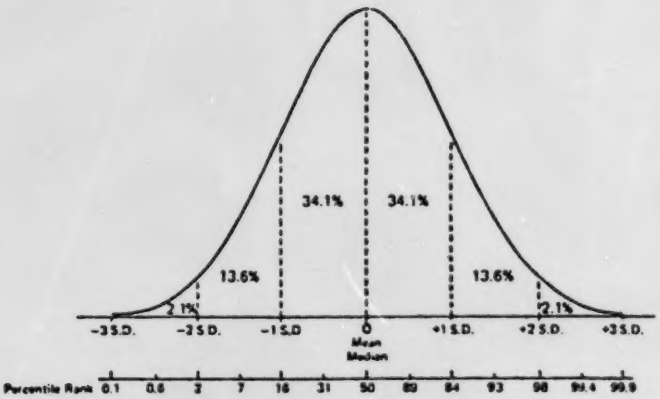
mastery test A test designed to determine whether a pupil has mastered a given unit of instruction or a single knowledge or skill; a test giving information on what a pupil knows, rather than on how his performance relates to that of some norm-reference group. Such tests are used in computer-assisted instruction, where their results are referred to as content - or criterion-referenced information.

non-verbal test A test that does not require the use of words in the item or in the response to it. (Oral directions may be included in the formulation of the task.) A test cannot, however, be classified as non-verbal simply because it does not require reading on the part of the examinee. The use of non-verbal tasks cannot completely eliminate the effect of culture.

normal distribution A distribution of scores or measures that in graphic form has a distinctive bell-shaped appearance. Figure 1 shows a graph of such a distribution. In such a normal distribution, scores or measures are distributed symmetrically about the mean, with as many cases up to various distances above the mean as down to equal distances below it. Cases are concentrated near the mean as down to equal distances below it. Cases are

concentrated near the mean and decrease in frequency, according to a precise mathematical equation, the farther one departs from the mean. The assumption that mental and psychological characteristics are distributed normally has been very useful in test development work.

Figure 1. Normal curve



norms Statistics that supply a frame of reference by which meaning may be given to obtained test scores. Norms are based upon the actual performance of pupils of various grades or ages in the standardization group for the test. Since they represent average or typical performance, they should not be regarded as standards or as universally desirable levels of attainment. The most common types of norms are deviation IQ, percentile rank, grade equivalent, and stanine. Reference groups are usually those of specified age or grade.

objective test A test made up of items for which correct responses may be set up in advance; scores are unaffected by the opinion or judgement of the scorer. Objective keys provide for scoring by clerks or by machine. Such a test is contrasted with a "subjective" test, such as the usual essay examination, to which different persons may assign different scores, ratings, or grades.

percentile (P) A point (score) in a distribution at or below which fall the percent of cases indicated by the percentile. Thus a score coinciding with the 35th percentile (P_{35}) is regarded as equaling or surpassing that of 35 percent of the persons in the group, and such that 65 percent of the performances exceed this score. "Percentile" has nothing to do with the percent of correct answers an examinee makes on a test.

percentile band An interpretation of a test score which takes account of the measurement error that is involved. The range of such bands, most useful in portraying significant differences in battery profiles, is usually from one standard error of measurement below the obtained score to one standard error of measurement above it.

percentile rank (PR) The expression of an obtained test score in terms of its position within a group of 100 scores; the percentile rank of a score is the percent of scores equal to or lower than the given score in its own or in some external reference group.

power test A test intended to measure level of performance unaffected by speed of response; hence one in which there is either no time limit or a very generous one. Items are usually arranged in order of increasing difficulty.

practice effect The influence of previous experience with a test on a later administration of the same or a similar test; usually an increased familiarity with the directions, kinds of questions, etc. Practice effect is greatest when the interval between testings is short, when the content of the two tests is identical or very similar, and when the initial test-taking represents a relatively novel experience for the subjects.

predictive validity See validity (2).

product-moment coefficient (r) Also known as the Pearson r. See coefficient of correlation.

profile A graphic representation of the results on several tests, for either an individual or a group, when the results have been expressed in some uniform or comparable terms (standard scores, percentile ranks, grade equivalents, etc.) The profile method of presentation permits identification of areas of strength or weakness.

prognosis (prognostic) test A test used to predict future success in a specific subject or field.

quartile One of three points that divide the cases in a distribution into four equal groups. The lower quartile (Q_1), or 25th percentile, sets off the lowest fourth of the group; the middle quartile (Q_2) is the same as the 50th percentile, or median, and divides the second fourth of cases from the third; and the third quartile (Q_3), or 75th percentile, sets off the top fourth.

r See coefficient of correlation.

range For some specified group, the difference between the highest and the lowest obtained score on a test; thus a very rough measure of spread or

raw score The first quantitative result obtained in scoring a test. Usually the number of right answers, number right minus some fraction of number wrong, time required for performance, number of errors, or similar direct, unconverted, uninterpreted measure.

readiness test A test that measures the extent to which an individual has achieved a degree of maturity or acquired certain skills or information needed for successfully undertaking some new learning activity. Thus a reading readiness test indicates whether a child has reached a developmental stage where he may profitably begin formal reading instruction. Readiness tests are classified as prognostic tests.

regression effect Tendency of a predicted score to be nearer to the mean of its distribution than the score from which it is predicted is to its mean. Because of the effects of regression, students making extremely high or extremely low scores on a test tend to make less extreme scores, i.e., closer to the mean, on a second administration of the same test or on some predicted measure.

reliability The extent to which a test is consistent in measuring whatever it does measure; dependability, stability, trustworthiness, relative freedom from errors of measurement. Reliability is usually expressed by some form of reliability coefficient or by the standard error of measurement derived from it.

reliability coefficient The coefficient of correlation between two forms of a test, between scores on two administrations of the same test, or between halves of a test, properly corrected. The three measure somewhat different aspects of reliability, but all are properly spoken of as reliability coefficients.

skewed distribution A distribution that departs from symmetry or balance around the mean, i.e., from normality. Scores pile up at one end and trail off at the other.

standard deviation (S.D.) A measure of the variability or dispersion of a distribution of scores. The more the scores cluster around the mean, the smaller the standard deviation. For a normal distribution, approximately two-thirds (68.3 percent) of the scores are within the range from one S.D. below the mean to one S.D. above the mean. Computation of the S.D. is based upon the square of the deviation of each score from the mean. The S.D. sometimes called "sigma" is represented by the symbol σ

standard error (S.E.) A statistic providing an estimate of the possible magnitude of "error" present in some obtained measure, whether (1) an individual score or (2) the mean of a group.

(1) standard error of measurement (SEM): As applied to a single obtained score, the amount by which the score may differ from the hypothetical true score due to errors of measurement. The larger the SEM, the less reliable the score. The SEM is an amount such that in about two-thirds of the cases the obtained score would not differ by more than one SEM from the true score. (Theoretically, then, it can be said that the chances are 2:1 that the actual score is within a band extending from true score minus 1 SEM to true score plus 1 SEM; but since the true score can never be known, actual practice must reverse the true-obtained relation for an interpretation.) Other probabilities are noted under (2) below. See **true score**.

(2) standard error: When applied to group average, standard deviations, correlation coefficients, etc., the S.E. provides an estimate of the "error" which may be involved. The group's size and the S.D. are the factors on which these standard errors are based. The same probability interpretation as for SEM is made for the S.E.'s of group measures, i.e., 2:1 (2 out of 3) for the 1 S.E. range, 19:1 (95 out of 100)

for a 2 S.E. range, 99:1 (99 out of 100) for a 2.6 S.E. range.

standard score A general term referring to any of a variety of "transformed" scores, in terms of which raw scores may be expressed for reasons of convenience, comparability, ease of interpretation, etc. The simplest type of standard score, known as a Z-score, is an expression of the deviation of a score from the mean score of the group in relation to the standard deviation of the scores of the group. Thus:

$$\text{standard score (Z)} = \frac{\text{raw score (X)} - \text{mean (M)}}{\text{standard deviation (S.D.)}}$$

Adjustments may be made in this ratio so that a system of standard scores having any desired mean and standard deviation may be set up. The use of such standard scores does not affect the relative standing of the individuals in the group or change the shape of the original distribution. T-scores have a M of 50 and a S.D. of 10. Deviation IQ's are standard scores with a M of 100 and some chosen S.D., most often 16; thus a raw score that is 1 S.D. above the M of its distribution would convert to a standard score (deviation IQ) of 100 + 16 = 116. (See Figure 1.)

Standard scores are useful in expressing the raw scores of two forms of a test in comparable terms in instances where try-outs have shown that the two forms are not identical in difficulty; also, successive levels of a test may be linked to form a continuous standard-score scale, making across-battery comparisons possible.

standardized test (standard test) A test designed to provide a systematic sample of individual performance, administered according to prescribed directions, scored in conformance with definite rules, and interpreted in reference to certain normative information. Some would further restrict the usage of the term "standardized" to those tests for which the items have been chosen on the basis of experimental evaluation, and for which data on reliability and validity are provided. Others would add "commercially published" and/or "for general use".

stanine One of the steps in a nine-point scale of standard scores. The stanine (short for standard-nine) scale has values from 1 - 9, with a mean of 5 and a standard deviation of 2. Each stanine (except 1 and 9) is 1/2 S.D. in width, with the middle (average) stanine of 5 extending from 1/4 S.D. below to 1/4 S.D. above the mean.

survey test A test that measures general achievement in a given area, usually with the connotation that the test is intended to assess group status, rather than to yield precise measures of individual performance.

test-retest reliability coefficient A type of reliability coefficient obtained by administering the same test a second time, after a short interval, and correlating the two sets of scores. "Same test" was originally understood to mean identical content, i.e., the same form; currently, however, the term "test-retest" is also used to describe the administration of different forms of the same test, in which case this reliability coefficient becomes the same

as the alternate-form coefficient. In either case (1) fluctuations over time and in testing situation, and (2) any effect of the first test upon the second are involved. When the time interval between the two testings is considerable, as several months, a test-retest reliability coefficient reflects not only the consistency of measurement provided by the test, but also the stability of the examinee trait being measured.

true score A score entirely free of error; hence, a hypothetical value that can never be obtained by testing, which always involves some measurement error. A "true" score may be thought of as the average score from an infinite number of measurements from the same or exactly equivalent tests, assuming no practice effect or change in the examinee during the testings. The standard deviation of this infinite number of "samplings" is known as the standard error of measurement.

validity The extent to which a test does the job for which it is used. Validity is always specific to the purposes for which the test is used. The term validity, then, has different connotations for various types of tests and, thus, a different kind of validity evidence is appropriate for each.

(1) content, curricular validity. For achievement tests, validity is the extent to which the content of the test represents a balanced and adequate sampling of the outcomes (knowledge, skills, etc.) of the course or instructional program it is intended to cover. It is best evidenced by a comparison of the test content with courses of study, instructional materials, and statements of education goals; and often by analysis of the processes required in making correct responses to the items. Face validity, referring to an observation of what a test appears to measure, is a non-technical type of evidence; apparent relevancy is, however, quite desirable.

(2) criterion-related validity. The extent to which scores on the test are in agreement with (concurrent validity) or predict (predictive validity) some given criterion measure. Predictive validity refers to the accuracy with which an aptitude, prognostic, or readiness test indicates future learning success in some area, as evidenced by correlations between scores on the test and future criterion measures of such success (e.g., the relation of scores on an academic aptitude test administered in high school to grade point average over four years of college). In concurrent validity, no significant time interval elapses between administration of the test to one, generally accepted as or known to be valid, or by the correlation between scores on a test and criteria measures which are valid but are less objective and more time-consuming to obtain than a test score would be.

(3) construct validity. The extent to which a test measures some relatively abstract psychological trait or construct; applicable in evaluating the validity of tests that have been constructed on the basis of an analysis (often factor analysis) of the nature of the trait and its manifestations. Tests of personality, verbal ability, mechanical aptitude, critical thinking, etc., are validated in terms of their construct and the relation of their scores to pertinent external data.

variability The spread or dispersion of test scores, best indicated by their standard deviation.

APPENDIX B

PERIODICALS THAT REVIEW STANDARDIZED TESTS

Educational and Psychological Measurement

3121 Cheek Road
Durham, North Carolina, U.S.A.
27708

Journal of Educational Measurement

National Council on Measurement in Education
206 South Kedzie Hall
Michigan State University
East Lansing, Michigan
44824

Measurement and Evaluation in Guidance

American Personnel and Guidance Association
1607 New Hampshire Avenue Northwest
Washington, D.C. 20009

Psychology in the Schools

4 Conant Street
Brandon, Vermont
05733

Titles available from: International Reading Association, Inc.
800 Barkadale Road
Newark, Delaware 19711

- Measuring Reading Performance. W.E. Blanton, R. Farr, J.J. Tuinman.
- Reading: What can be Measured? R. Farr.
- Reading Tests for the Secondary Grades. W. Blanton, R. Farr, J.J. Tuinman.
- Tests of Reading Readiness and Achievement. R. Farr and N. Anastasiou.
- How to use W.I.S.C. Scores in Reading Diagnosis. E.F. Searls.

APPENDIX C

A SELECTIVE LIST OF TEST PUBLISHERS

Addison-Wesley (Canada) Ltd.
P.O. Box 580
36 Prince Andrew Place
Don Mills, Ontario M3C 2T8

CTB/McGraw-Hill Ryerson Ltd.
330 Progress Avenue
Scarborough 707, Ontario

Guidance Centre
Faculty of Education
University of Toronto
1000 Younge Street
Suite #304
Toronto, Ontario M4W 2K8

Psycan Ltd.
255 Consumers Road
Willowdale, Ontario
M2J 1R3

SRA
P.O. Box 707
Station 'A'
Willowdale, Ontario
M2N 5T5

The Psychological Corporation
Polk & Geary
San Francisco, California
94109

Thomas Nelson & Sons (Canada) Ltd.
81 Curlew Drive
Don Mills, Ontario M3A 2R1

APPENDIX D

STANDARDIZED SCHOLASTIC APTITUDE TESTS

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
Analysis of Learning Potential (1970)	1-12		Reading, mathematics, complete battery	Guidance Centre
Canadian Lorge-Thorndike (1967)	3-9		Verbal, non-verbal, composite	Thomas Nelson & Sons
Cooperative School and College Ability Tests, Series 2 (1973)	3-14	A,B	Verbal, quantitative, total	Educational Testing Service, Addison-Wesley
Differential Aptitude Tests (1974)	8-12	S,T	Verbal reasoning, numerial ability, abstract reasoning, clerical speed and accuracy, mechanical reasoning, space relations, spelling, language usage	Guidance Centre
Henmon-Nelson Tests of Mental Ability (1973)	3-12		Total score only	Thomas Nelson & Sons

STANDARDIZED SCHOLASTIC APTITUDE TESTS (CONTINUED)

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
Otis-Lennon Mental Ability Test (1967)	K-Adult	J,K	One score converted to various derived scores in age and grade categories	Guidance Centre

STANDARDIZED ACHIEVEMENT BATTERIES

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
California Achievement Test (1970)	1.5-2 2-4 4-6 6-9 9-12	A,B	Language (Auding, Mechanics, Usage and Structure, Spelling) Reading (Vocabulary, Comprehension), Mathematics (Computation, Concepts and Problems), Language (Mechanics, Usage, Structure, Spelling) (Same subtests used on the four highest levels)	CTB/McGraw-Hill Ryerson
Canadian Tests of Basic Skills (1976)	1-2 2-3	7M 8M	Vocabulary, Word Analysis, Reading, Spelling and Mathematics Problems and Concepts	Thomas Nelson and Sons
	3 4 5 6 7 8	3M 4M	Vocabulary, Reading, Spelling, Capitalization, Punctuation, Usage, Map Reading, Reading Graphs and Tables, Use of Reference Material, Concepts, and Problems (Same subtests used throughout levels 3-8)	

STANDARDIZED ACHIEVEMENT BATTERIES (CONTINUED)

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
Comprehensive Tests of Basic Skills (1973)	K.0-1.3	S	Readiness	CTB/McGraw-Hill Ryerson
	K.6-1.9	S	Reading, Language, Mathematics	
	1.6-2.9	S	Reading (Vocabulary, Comprehension),	
	2.5-4.9	S,T	Language (Mechanics,	
	4.5-6.9	S,T	Expression, Spelling),	
	6.5-8.9	S,T	Mathematics (Computation,	
	8.5-12.9	S,T	Concepts, Applications), Study Skills (Reference Materials, Graphic Materials) (Same subjects used throughout levels)	
Iowa Tests of Educational Development (1972)	9-12	X5, Y5	Reading, Language Arts, Mathematics, Social Studies, Science and Uses of Sources	SRA
Metropolitan Achievement Test (1970)	K.7-1.4	F,H	Listening for Sounds, Reading, Numbers	Guidance Centre
	1.5-2.4	F,G,H	Word Knowledge, Word Analysis, Reading, Mathematics Computation, Mathematical Concepts	
	2.5-3.4	F,G,H	Word Knowledge, Word Analysis, Reading, Spelling, Mathematics Computation, Mathematical Concepts, Mathematics Problem Solving	
	3.5-4.9	F,G,H	Word Knowledge, Reading, Language, Spelling, Mathematics Computation, Mathematical Concepts, Mathematics Problem Solving	

STANDARDIZED ACHIEVEMENT BATTERIES (CONTINUED)

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
Metropolitan Achievement Test (1970) (continued)	5.0-6.9	F,G,H	Word Knowledge, Reading, Language, Spelling, Mathematics Computation, Mathematical Concepts, Mathematics Problem Solving, Science, Social Studies	
	7.0-9.5	F,G,H	Same as above	
SRA Achievement Series (1971)	1-3	E,F	Reading (Word Verbal-picture Association, Sentence picture Association, Comprehension, Vocabulary), Mathematics (Concepts and Computation), Language Arts (Alphabetization, Capitalization, Punctuation, Spelling and Usage)	SRA
	2-4	E,F	Reading (Word-Verbal-picture Association, Sentence-picture Association, Comprehension, Vocabulary), Mathematics (Concepts and Computation), Language Arts (Alphabetization, Capitalization, Punctuation, Spelling and Usage)	
	4-5	E,F	Reading (Comprehension, Vocabulary, and Total), Language Arts (Usage, Spelling and Total), Mathematics (Concepts, Computation, and Total), Social Studies, Science, Uses of Sources	
	6-7	E,F	Same as above	
	8-9	E,F	Same as above	

STANDARDIZED ACHIEVEMENT BATTERIES (CONTINUED)

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
Sequential Tests of Educational Progress Series II (STEP) (1969)	4-6	A,B	English Expression, Reading, Mechanics of Writing, Mathematics Computation, Mathematics Basic Concepts, Science, Social Studies (Mechanics of Writing, Mathematics Basic Concepts for grade 4-12 only)	Educational Testing Service
	7-9	A,B		
	10-12	A,B		
	13-14	A,B		
Stanford Achievement Test (1973)	1.5-2.4	A,B	Vocabulary, Reading Comprehension, Word Study Skills, Mathematics Concepts, Mathematics Computation, Spelling, Listening Comprehension	Guidance Centre
	2.5-3.4	A,B	Vocabulary, Reading Comprehension, Word Study Skills, Mathematics Concepts, Mathematics Computation, Mathematics Application, Spelling, Social Studies, Science, Listening Comprehension	
	3.5-4.4	A,B	Vocabulary, Reading Comprehension, Word Study Skills, Mathematics Concepts, Mathematics Computation, Mathematics Application, Spelling, Language, Social Science, Science, Listening Comprehension	
	4.5-5.4	A,B	Same as above	
	5.5-6.9	A,B	Same as above	
	7. -9.5	A,B	Vocabulary, Reading Comprehension, Mathematics Concepts, Mathematics Computation, Mathematics Application, Spelling, Language, Social Science, Science	

STANDARDIZED ACHIEVEMENT BATTERIES (CONTINUED)

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
Stanford Early School Achievement Test (1970)	K.1-1.1		Environment, Mathematics, Letters and Sounds, Aural Comprehension	Guidance Centre
	1.1-1.8		Environment, Mathematics, Letters and Sounds, Aural Comprehension, Word Reading and Sentence Reading	
Stanford Test of Academic Skills (1973)	8-10	A,B	Reading, English & Mathematics	Guidance Centre
	11-12	A,B	Same as above	
	College	A,B	Same as above	
Tests of Academic Progress (1972)	9-12	S	Social Studies, Composition, Science, Reading, Mathematics and Literature	Thomas Nelson & Sons

STANDARDIZED INTEREST AND VOCATIONAL TESTS

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
Vocational Planning Inventory	8-10 11-13	I, 2 levels	SRA Arithmetic Index, SRA Pictorial Reasoning Test, SRA Verbal Form, Survey of Interpersonal Values, Survey of Personal Values, Mechanics subtest of Flanagan Aptitude Classification Tests, Flanagan Industrial Tests: arithmetic, assembly, expression, memory, scales, tables.	SRA

STANDARDIZED INTEREST AND VOCATIONAL TESTS (CONTINUED)

Test	Grade Levels	Forms	Subtests	Publisher or Distributor
California Occupational Preference Survey	9-16	I	Science professional, science skilled, technical professional, technical skilled, outdoor, business professional, business skilled, clerical, linguistic professional, linguistic skilled, aesthetic professional, aesthetic skilled, service professional, service skilled	Educational and Industrial Testing Service
Kuder General Interest Survey	6-12	I	Outdoor, mechanical, computational, scientific, persuasive, artistic, literary, musical, social service, clerical, verification	SRA
Safran Student's Interest Inventory	8-12	I	Interest Scores: economic, technical, outdoor, service, humane, artistic, scientific Ability Self-ratings: academic, mechanical, social, clerical	Thomas Nelson & Sons
Strong Vocational Interest Blank	Age 16 & over		22 basic interests; 54 occupational; 8 non-occupational	Stanford University Press

APPENDIX E

TEST ADMINISTRATION GUIDE

BEFORE THE TESTING DATE

1. Understand nature and purposes of the testing
   a) Tests to be given
   b) Reasons for giving tests
2. Decide on number to be tested at one time
3. Decide on seating arrangements
4. Decide on exact time of testing
   a) Avoid day before holiday
   b) Avoid conflicts with recess of other groups
   c) Make sure there is ample time
5. Procure and check test materials
   a) Directions for administering
   b) Directions for scoring
   c) Test booklets
      - one for each pupil and examiner
   d) Answer sheets
      - one for each pupil and examiner
   e) Pencils (regular or special)
   f) Stopwatch or other suitable timer
   g) Scoring keys
   h) "Testing - Do Not Disturb" sign
   i) Other supplies (scratch paper, etc.)
6. Study test and directions carefully
   a) Familiarize yourself with:
      - general make-up of test
      - time limits
      - directions
      - method of indicating answers
   b) Take the test yourself

turn to page 76 for three remaining points.

JUST BEFORE TESTING

1. Put up "Testing - Do Not Disturb" sign
2. See that desks are cleared
3. See that pupils have sharpened pencils
4. Attend to toilet needs of pupils
5. Check lighting
6. Check ventilation
7. Make seating arrangements

DURING TESTING

1. Distribute materials according to predetermined order
2. Caution pupils not to begin until you tell them to do so
3. Make sure that all identifying information is written on booklet or answer sheet
4. Read directions exactly as given
5. Give signal to start
6. Write starting and finishing times on the chalkboard
7. Move quietly about the room to:
   a) make sure pupils are marking answers in the correct place
   b) make sure pupils are continuing to the next page after finishing the previous page
   c) make sure pupils stop at the end of the test
   d) replace broken pencils
   e) encourage pupils to keep working until time is called
   f) make sure there is no copying

DURING THE TESTING is continued on page 76.

BEFORE THE TESTING DATE (continued)

7. Arrange materials for distribution

   a) Count number needed

8. Decide on order in which materials are to distributed and collected

9. Decide what pupils who finish early are to do

DURING TESTING (continued)

   g) at end to pupils finishing early

8. Permit no outside interruptions

9. Stop at the proper time

JUST AFTER TESTING

1. Collect materials according to predetermined order

2. Count booklets and answer sheets

3. Make a record of any incidents observed that may tend to invalidate scores made by pupils

APPENDIX F

SELECTING AND USING A STANDARDIZED TEST: PROCEDURES CHECKLIST

REFERENCES

1. For what purpose(s) is this test being used?

2.1(a) Selection? (__) Explain: _____

_____

2.1(b) Placement? (__) Explain: _____

_____

2.1(c) Diagnosis/Remediation? (__) Explain: _____

_____

2.1(d) Program Planning and Evaluation? (__) Explain: __

_____

2. List several tests which could be considered for your purpose(s):

1.3
Appendices B, C, D
P.33
6.3

1. _____
2. _____
3. _____
4. _____

3. Rate each test on the following criteria:

1.3	1. Buros' review of each test	1 2 3 4 5
2.2(a)	2. Content of test is relevant	1 2 3 4 5
2.2(b)(c)	3. Available norms are appropriate and up-to-date	1 2 3 4 5
3.2,3	4. Costs in purchase, administration time, and scoring are satisfactory	1 2 3 4 5
4.1,2	5. Potential for long-term use is high	1 2 3 4 5
4.1,2,3	6. Interpretation of results by all interested parties is straight-forward	1 2 3 4 5

(Sum the ratings for each test on these criteria and others you may deem important. Select the test with the highest combined rating.)

SELECTING AND USING A STANDARDIZED TEST:
PROCEDURES CHECKLIST (CONTINUED).

REFERENCES

4. What other individuals and groups have an interest in the outcome of the testing?

2.1
5.1,2,3

1. Teachers? (__) _____
2. Pupils? (__) _____
3. Parents? (__) _____
4. Trustees? (__) _____
5. Public? (__) _____
6. Consultants? (__) _____
7. Principals? (__) _____
8. Other Administrators? (__) _____
9. Support Staff (__) _____

2.2
3
4
5

5. Specify how and to what extent each of the interested groups should be involved in (a) selecting, (b) administering, (c) scoring, and (d) interpreting of the test.

	TESTING STAGES			
INTERESTED GROUPS	SELECTION	ADMINISTERING	SCORING	INTERPRETING
TEACHERS				
PUPILS				
PARENTS				
TRUSTEES				
PUBLIC				
CONSULTANTS				
PRINCIPALS				
OTHER ADMINISTRATORS				
SUPPORT STAFF				
OTHER				

(1) check those individuals or groups with a role to play in each stage of the testing process

(2) specify what tasks constitute each group's function at each stage selected.

SELECTING AND USING A STANDARDIZED TEST:
PROCEDURES CHECKLIST (CONTINUED).

REFERENCES

6. Who will make recommendations for actions as a result of the testing, and how will these be carried out?

7. Decide how the results will be communicated to each of the groups checked in Section 4.

5.1,2,3

APPENDIX G

DISTRICT LEVEL RESOURCES ESSENTIAL TO A VIABLE TESTING PROGRAMME

RESOURCE	PURPOSES
Tests in Print II. (Ed.) O.K. Buros, The Gryphon Press, Highland Park, N.J., 1974	- indexes over 16,000 tests with references. - surveys available published tests under any given topic.
The Eighth Mental Measurement Yearbook, Volumes I and II, (Ed.) O.K. Buros, The Gryphon Press, Highland Park, N.J., 1978.	- describes and references published tests in terms of: client group, subtests, scoring options, costs, and publisher. For many tests, one or more objective reviews of the technical strengths and weaknesses of the tests are included.
Specimen Sets of Tests	- available from publishers free or at nominal costs. Contains sample of test, description by publisher, norms, technical manual, directions for administration and scoring. - necessary for validity reasons after initial screening of possibilities, especially for achievement tests.
French, J.W., and Michael, W.B. Standards for Psychological Tests and Manuals. Washington: American Psychological Association, 1974.	- describes standards expected of published tests. - provides criteria against which possible choices of tests can be compared.
Cronbach, L.J. Essentials of Psychological Testing (Second Edition). New York: Harper and Row, 1960 OR Anastasi, A. Psychological Testing (Fourth Edition). New York: Macmillan, 1976.	- classic texts on the purposes, uses, and interpretation of educational and psychological testing.